2015

Spectrum Allocation in Networks with Finite Sources and Data-Driven Characterization of Users' Stochastic Dynamics

Ahsan-Abbas Ali
Louisiana State University and Agricultural and Mechanical College

SPECTRUM ALLOCATION IN NETWORKS WITH FINITE SOURCES AND DATA-DRIVEN CHARACTERIZATION OF USERS' STOCHASTIC DYNAMICS

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Electrical and Computer Engineering

by
Ahsan-Abbas Ali
B.Sc., University of Engineering and Technology at Lahore, 2007
August 2015

To my family:

Nauroze Tahira, Ghazanfar Ali, Mehak, Muqadas, Hira, Shan, Mehdi, Alamdar.

Acknowledgements

Firstly, I would like to express my sincerest gratitude to my major advisor, Prof. Shuangqing Wei. I am grateful to him for his consistent and invaluable support and guidance, which he provided me throughout my tenure as a PhD student at LSU. The beginning and completion of this long journey, full of challenges, would not have been possible without his generous support and efforts. I would also like to take this opportunity to extend my sincerest thanks to Prof. Jian Zhang for serving as a member on my advisory committee, and also for his invaluable time and efforts that he spent in collaborating with us in our projects. I want to express my special gratitude to all the other respected members of my advisory committee, Prof. Padmanabhan Sundar, Prof. Xue-Bin Liang, Prof. Ramachandran Vaidyanathan, and Prof. Cassandra Chaney for their generous support and cooperation. Also, I would like to thank all my teachers of EECS and Mathematics for a great learning experience that I had during my graduate coursework at LSU. Finally, I am obliged to my colleagues and friends who supported me during my PhD years. I thank Muhammad Anis-ur-Rehman, Yahya Khiabani, Phuoc Vu, Shuhang Wu and Ali Moharrer for their warm friendship and support.

Table of Contents

List of Tables

List of Figures

Abstract

During emergency situations, the public safety communication systems (PSCSs) get overloaded with high traffic loads. Note that these PSCSs are finite source networks. The goal of our study is to propose techniques for an efficient allocation of spectrum in finite source networks that can help alleviate the overloading of PSCSs.

In a PSCS, there are two system segments, one for the system-access control and the other for communications, each having dedicated frequency channels. The first part of our research, consisting of three projects, is based on modeling and analysis of finite source systems for optimal spectrum allocation, for both access-control and communications. In the first project, Chapter 2, we study the allocation of spectrum based on the concept of cognitive radio systems. In the second project, Chapter 3, we study the optimal communication channel allocation by call admission and preemption control. In the third project, Chapter 4, we study the optimal joint allocation of frequency channels for access-control and communications.

Note that the aforementioned spectrum allocation techniques require the knowledge of the call traffic parameters and the priority levels of the users in the system. For practical systems, these required pieces of information are extracted from the call records meta-data. A key fact that should be considered while analyzing the call records is that the call arrival traffic and the users priority levels change with a change in events on the ground. This is so because a change in events on the ground affects the communication behavior of the users in the system, which affects the call arrival traffic and the priority levels of the users. Thus, the first and the foremost step in analyzing the call records data for a given user, for extracting the call traffic information, is to segment the data into time intervals of homogeneous or stationary communication behavior of the user. Note that such a segmentation of the data of a practical PSCS is the goal of our fourth project, Chapter 5, which constitutes the second part of our study.

Chapter 1
Introduction

1.1 Motivation: Public Safety

The main motivation of this study is the overloading issue of the *public safety communication systems* (PSCSs) during emergency situations. In these situations, an overwhelmingly large amount of traffic pours into the network over a hot spot area. This poses a serious threat to the quality of communications among the first responders, given the limited and fixed amount of network resources to dispense with. Eventually it further creates a threat to public safety and the well-being of civilian life. Note that PSCSs are *finite source* networks, as described in Section 1.2. Thus, the goal of this study is to propose techniques for an efficient allocation of spectral resources in such systems. These techniques help alleviate the overloading of PSCSs, which in turn ensures the required quality of communications among the first responders during emergency situations. A detailed summary of the study is provided in Section 1.3.

1.2 Public Safety Communications: Finite Source Networks

In PSCSs, the coverage area is divided into *sites*. Also, the users are divided into fleets or groups normally called *talk groups*. Each user can only communicate with another user of the same talk group. When a user needs to talk, it presses the push-to-talk (PTT) button of its radio, in order to send a call request at the site's base station or *access-station* (AS), over a *control channel* which is dedicated for the access control process. The AS then assigns the communication resources to the caller and the rest of the users of the talk group, in order to broadcast the voice call throughout the group, as shown in Figure 1.1. In this figure, a single site with multiple talk groups is shown. Note that when the radios from the same talk-group are dispersed over multiple sites, a call from the group cannot be established until all the radios of the group are reachable with resources to serve in associated sites. Such a scenario is shown in Figure 1.2.

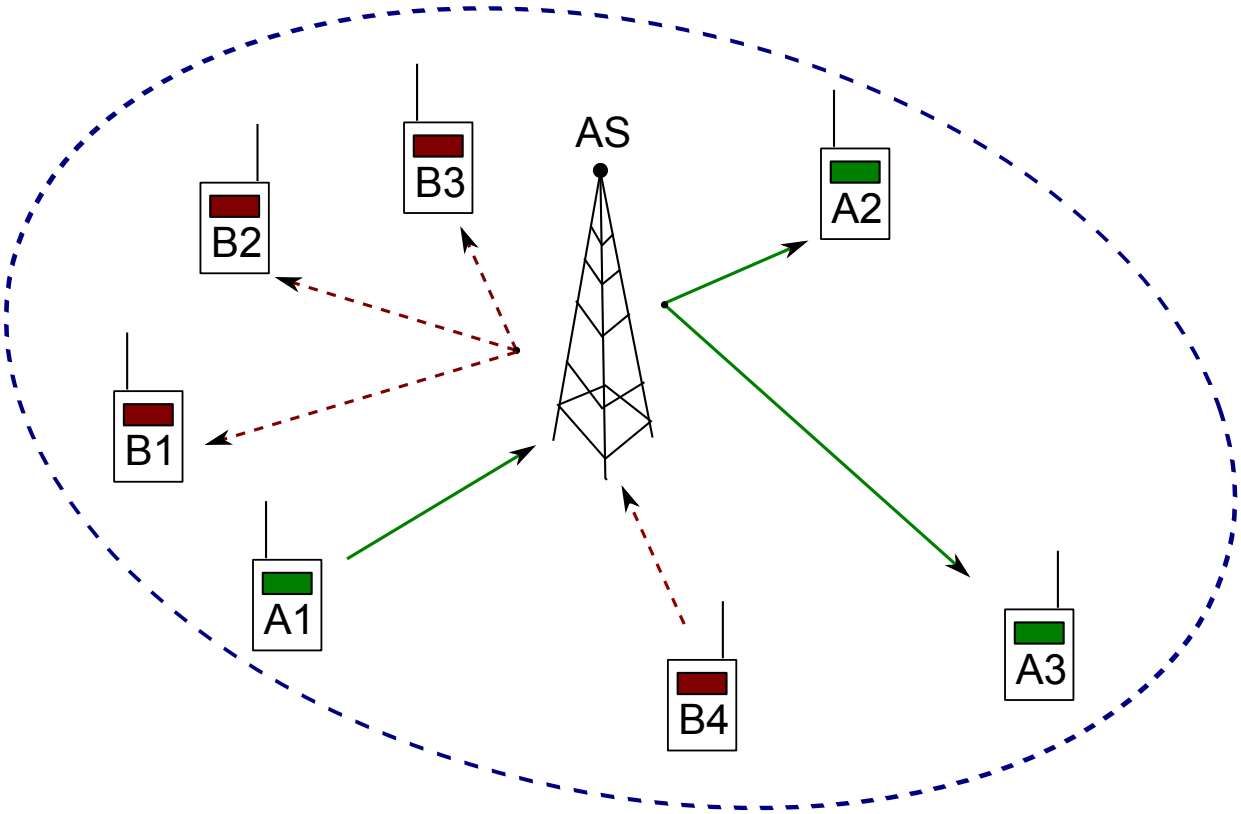


FIGURE 1.1: PSCS with two talk groups, A and B, within the same site.

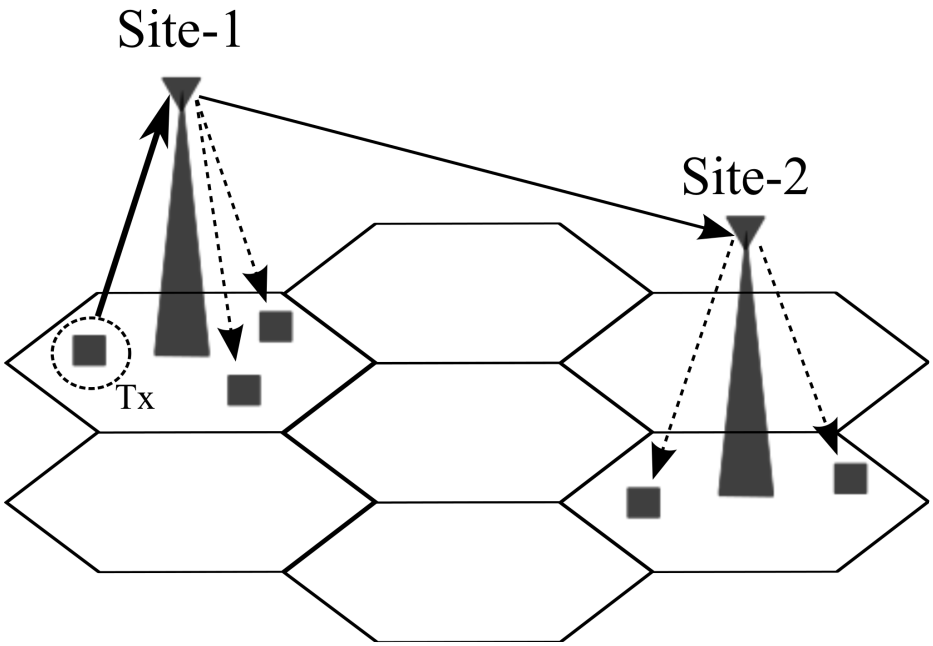


FIGURE 1.2: PSCS with a group dispersed over multiple sites. Tx is a transmitter or caller radio.

Thus, in PSCSs, the channel allocation for communications is performed on the basis of talk groups. Therefore, instead of radios, the call source units are the talk groups. Note that the PSCSs have the following properties of the *finite source* systems:

1. The call source units, i.e., the talk groups, are finite in number.

2. When a user in a talk group talks, the rest of the users only listen. Therefore, a talk-group under service cannot generate any call. Thus, only an idle source unit, i.e., an idle talk-group can generate a call in a PSCS.

1.3 Summary of Research Work

We divide our study into two parts as follows:

1. *Part-I:* System Modeling and Analysis for Spectrum Allocation.

2. *Part-II:* Data-Driven Characterization of Users' Stochastic Dynamics.

We briefly discuss both these parts of our study in the subsequent sub-sections.

1.3.1 Part-I: System Modeling and Analysis for Spectrum Allocation

In a control channel based wireless access system, like the one discussed in Section 1.2, there are two types of frequency channels. One of the types is the access-control or simply *control channel* that is dedicated for the access control process, whereas the rest of them are of the type of *communication channels*. The first part of our research is based on the modeling and analysis of the finite source systems for the optimal allocation of frequency channels, for both access-control and communications. We accomplish this goal in the form of following three projects:

1. Efficient spectrum allocation via cognitive radio based system design.

2. Communication channel allocation by call admission and preemption control.

3. Optimal joint allocation of control and communication channels.

The first project is discussed in Chapter 2. In this project, we study how an efficient spectrum allocation can be achieved by adopting a *cognitive radio* based design approach. Note that only this project deals with the physical (PHY) layer of the communication system, whereas the second and third projects deal with the medium access control (MAC) layer. In a cognitive radio based system, there are two types of users, namely, primary users (PUs) and secondary users (SUs). PU is a high priority user that can use the allocated frequency channel at any time when needed. On the other hand, SU is a low priority user that can use the channel allocated for PU, when PU is idle and the channel is available for SU. One of the challenges faced while designing such systems is to improve the capability of the SU transmitter (SU-TX) to detect the availability of the channel. This can be achieved by appropriately selecting the detection parameters of SU-TX's signal detector that result in an optimal *sensing-throughput tradeoff*, as explained in Chapter 2. In that chapter, we provide the design guidelines for the cognitive radio systems, based on the transmission-power levels of the users. Note that the design goal in this project is to maximize the SU performance by keeping the PU performance degradation within a given tolerable range.

The second project is discussed in Chapter 3. In this project, we study the communication channel allocation as the call admission and preemption control, for finite source systems. To this end, the system is modeled as a prioritized queueing system, wherein the users are assigned priority levels, and the system operation is modeled as a Markov decision process (MDP). Based on these priorities, optimal decisions are determined for admitting a new call or preempting an already busy call. These policy decisions depend on the priority level of the call and the state of the system. In this study, we demonstrate that the optimal policies are *threshold based* policies. Note that the threshold based policies are easy to design and can be implemented in terms of decision thresholds. These decision thresholds are used to make optimal admission and preemption decisions during the system operation.

We discuss the third project in Chapter 4. In that chapter, we consider a control channel based wireless access system, like the one discussed in Section 1.2. In such systems there are two types

of frequency channels. One of the types is called the control channel that is dedicated for access control, whereas the rest of the channels are used for communications. Therefore, we can divide the system into two segments or layers, namely, the access layer and the communication layer. In order to implement a bandwidth efficient system we need an optimal allocation of control and communication channels. As explained in detail in Chapter 4, the performance of one layer significantly affects that of the other one, in case of the finite source wireless systems like PSCSs. Thus, for such systems, the control and communication layers are inseparable. Therefore we need a system model that can model both the layers jointly, and thus, can help us devise a mechanism to optimally allocate the control and communication channels in a wireless access system. However, in conventional studies on wireless access systems, these layers are modeled separately. Note that these separate control and communication layer models are only applicable under certain assumptions that are discussed in Chapter 4, and do not hold for the finite source systems like PSCSs. Therefore, in this study, we propose a novel statistical model for wireless access systems that jointly models the control and communication layers, and helps evaluate the optimal number of control and communication channels. We also propose the concept of a *channel allocation map* that helps visualize the optimal channel allocation for all possible values of the call traffic parameters. Note that the optimal channel allocation also requires the knowledge of the actual call-arrival traffic load. However, this load is invisible to a practical system, because in practice, a system does not keep records of the calls that are blocked due to collision at the access layer. Therefore, in Chapter 4, we also demonstrate the capability of our proposed model in estimating the invisible actual traffic load. Finally, we provide guidelines for developing an algorithm for the traffic aware allocation of channels, based on our proposed model.

1.3.2 Part-II: Data-Driven Characterization of Users' Stochastic Dynamics

The spectrum allocation techniques, e.g., the ones discussed in Part-I of this study, requires the knowledge of the call traffic parameters and the priority levels of the users in the system. For practical systems, these required pieces of information are extracted from the call records meta-

data. A key fact that should be considered while analyzing the call records is that the call arrival traffic and the users priority levels change with a change in events on the ground. This is so because a change in events on the ground affects the communication behavior of the talk groups in the system, which affects the call arrival traffic and the priority levels of the users. For example, a change from a normal scenario to an emergency situation, which causes a change in the talk groups' activities of a PSCS. Thus, the first and the foremost step in analyzing the call records data for a given talk group, for extracting the call traffic information, is to segment the data into time intervals of homogeneous or stationary communication behavior of the group. Note that such a segmentation of the data of a practical PSCS is the goal of Part-II of our study.

This part of the study consists of only a single project discussed in Chapter 5. In this project, we use the call traffic meta-data of the State of Louisiana's PSCS. We develop a way to quantify a talk group's activity as a discrete sequence of symbols. Here a symbol, corresponding to a call, represents the calling radio-unit in the talk group, as explained in Chapter 5. Note that we extract this talk group activity from the available data of call records. As described in the last paragraph, in practice, a talk group's communication behavior remains consistent during a certain event happening on the ground. However, with a change in event, the talk group's behavior also changes. Thus, a talk group's activity, extracted as a discrete sequence, from the whole day data of call records, is not stationary. However, it may consist of many stationary segments depending on different events that occurred on the ground. Our primary goal is to segment a given whole day activity of a talk group into stationary segments, each corresponding to a distinct event, and also quantify the behavior of the talk group within each segment or event. Note that we quantify the behavior of a talk group in the form of a *context tree* which represents a *variable memory Markov process* of discrete symbols, as explained in Chapter 5. For this project, we use the data mining techniques that have demonstrated state-of-the-art results in diverse fields, e.g., bioinformatics (gene sequence analysis), linguistics (language modeling) and information theory (data compression). We have also made non-trivial modifications to these techniques to match our requirements.

In Chapter 5, we discuss in detail the representation of a talk group's communication activity in terms of a sequence of discrete symbols. We also provide the details of the algorithm that we have developed for the segmentation of discrete sequences. Finally, we present the segmentation results for the synthetic and the real data sets that demonstrate the validity of the proposed algorithm.

Chapter 2

Spectrum Allocation via Cognitive Radio Based Systems

The users of sophisticated broadband wireless services are increasing day by day, thereby raising the demand of these services. Therefore, already established networks are expanding their resources and the new service providers are establishing their infrastructure. This gives rise to the problem of scarce bandwidth resources. Since most of the available spectrum has already been licensed, there is almost no room left for accommodating the new demands. There are studies like [1] which show that vast regions of licensed spectrum are underutilized. These are called *white-spaces*. A prospective solution to the problem of scarce bandwidth resources is to use these white spaces for the new wireless systems instead of issuing the new licenses. The device that can help us achieve this goal of utilizing the unused channel is the *cognitive radio* [2, 3]. It is a radio that can sense and learn, as the word *'cognitive'* indicates. It has intelligent capabilities to sense the communication activities over the channel and looks for the opportunities available for itself.

During emergency situations where multiple public safety talk groups are to share resources due to the lack of sufficient spectrum to accommodate all surging needs, spectrum sharing becomes necessary. Cognitive radio based systems are proposed to enable such spectrum sharing among the first responder groups during emergency situations [4–7]. In this study we consider a cognitive radio (CR) based prioritized public safety communication system. In this system, there are two categories of talk groups or users. One of them is the high priority talk group called the *primary user* (PU), and the other one is the low priority talk group called the *secondary user* (SU). PU's are the users which possess the license for using the channel and can use it for communications at anytime. On the other hand, SU's are the unlicensed users and can only use the channel for communications when it is idle. These SU's form the nodes of the cognitive radio network and are intelligent enough to sense whether the channel is being used by a PU or not. This is referred

to as the *channel sensing* ability of the SU. Also, SU should leave an occupied channel whenever a PU starts transmission using that channel.

In this study, we propose a mixture-Gaussian model for a cognitive radio channel to analyze the interplay between the interference in the system and the degrees-of-freedom (DOF), i.e., the number of channel uses, used by the secondary user (SU) for communications. In contrast to the conventional studies, we assume that the SU receiver (SU-RX) does not precisely know whether the primary-user (PU) transmitter is on or off. Due to this assumption the resulting interference channel is mixture-Gaussian. Our objective is to find the optimal sensing threshold and sensing time for the signal detector used by the SU transmitter (SU-TX). Our formulation of the optimization problem reflects the trade-off between SU-TX's DOF for communications and that for detection. Both the DOFs affect PU's interference to SU, and SU's interference to PU. The latter interference causes PU performance degradation, which is kept within tolerable range as a constraint. As a further contribution, we define interference regimes for SU performance on the basis of PU transmission power level. We also address the scenario when PU receiver uses the nearest neighbor decoding while wrongly anticipating that the channel is Gaussian. Finally, we demonstrate that even if SU-TX's signal detector performs suboptimally, SU can still achieve the optimal detector's performance in the high interference regime by adjusting the sensing parameters. This study has resulted in two research papers, [8, 9].

2.1 Introduction

We consider a simple cognitive radio system consisting of two transmitter-receiver pairs, one for the primary user PU and the other for the secondary user SU. They have a common frequency channel to use for communications and SU communicates whenever the channel is sensed idle. In order to find out whether the channel is occupied by a PU signal or not, SU transmitter uses a signal detector, e.g., the energy detector. For this system, the main design problem is to find the optimal sensing time and threshold for the detector used by SU transmitter. Here we would like to mention an important remark, i.e., for the rest of the discussion, the *signal detector of SU* refers

to the *signal detector used by the SU transmitter*. The optimality criterion is to maximize the SU performance while keeping the detection probability above an appropriate threshold level. This criterion is also called the *sensing-throughput trade-off* for SU.

Note that in this study, *coherence* means the availability of the channel state information (CSI) to the receiver [10–12]. In our model, CSI for the SU receiver is the precise knowledge of the PU transmitter's transmission state, i.e., whether it is on or off. The sensing-throughput trade-off and analysis models have been explored in [13–26], under an assumption that SU receiver operates coherently with both SU and PU transmitters, i.e., SU receiver precisely knows if SU and PU transmitters are on or off. We call this the *conventional model*. However, in this study, we consider a more practical scenario where SU receiver operates coherently with SU transmitter but *incoherently* with PU transmitter, as in our previous work [27]. This means that SU receiver precisely knows if SU transmitter is on or off, but it does not know the same about PU transmitter. This model appropriately incorporates the interference experienced by both PU and SU systems and is practically more rigorous than the conventional model. To the best of our knowledge, despite such significance, this model has never been studied.

The assumption that whether the SU receiver operates coherently or incoherently with the PU transmitter, determines the nature of the interference channel in the system model. In the conventional model, due to the coherent operation of SU receiver and PU transmitter, the underlying channel is a regular Gaussian interference channel. On the other hand, in our model, due to the incoherent operation of SU receiver and PU transmitter, the resulting interference channel is mixture-Gaussian, as explained in Section 3.2. Due to the mixture-Gaussian nature of the interference channel, the optimal sensing time and threshold design problem can be better elaborated with the help of a novel concept of a trade-off, that we propose formally in Section 2.5.1 as the trade-off between the degrees of freedom available to SU for communications and the interference experienced by SU due to PU, rather than that of the conventional sensing-throughput trade-off. In this study, the degrees of freedom available to SU for communications quantify the average

number of channel-uses per transmission frame, which is used by SU for communications in the long run, i.e., for infinitely large number of transmission frames, as explained in detail in Section 2.5.1. We properly define and elaborate the novel concept of the trade-off in Section 2.5.1, and before proceeding to it, we also build the required background in the subsequent sections, in order to help the reader understand the concept nicely.

Due to the incoherence assumption between SU receiver and PU transmitter, the analysis becomes more challenging. The first challenge is that the interference channel becomes mixture Gaussian that requires the evaluation of the entropy of mixture Gaussian random variables, which does not have a closed form solution. The second challenge is the non-linear interdependence of quite a few parameters. Thus, the optimization problem under consideration becomes quite complicated and cannot be solved using the conventional optimization techniques, as explained in Section 2.3.2. Also, it does not have a closed form solution, and therefore, we develop numerical algorithms to solve it.

Another loose end in the conventional studies is that an arbitrarily high detection-probability threshold is selected for the detection-probability constraint. In order to tie up this loose end, we select a meaningful value for this threshold which ensures that the PU performance degradation remains within a tolerable range. This requires us to explicitly compute PU performance. While evaluating PU performance, we assume that PU receiver does not know the state of SU transmitter, i.e., PU receiver also operates incoherently with SU transmitter, in the same way as we assume that SU receiver operates incoherently with PU transmitter. Note that a degradation in PU performance is caused by the interference due to SU, when the detection probability is less than 1.

As a further novel contribution, we identify a performance inefficiency region for SU at a very high detection probability, where SU performance drastically decreases with increase in detection probability. We call this the *energy detector's inefficiency region* and is abbreviated as EDI region. We also identify the interference regimes for SU performance on the basis of PU transmission power level $\mathcal{P}_{\mathrm{PU}}$. Our study reveals that the tolerance level of PU performance can be exploited to

increase the SU performance by sacrificing the detection probability, but only for the *low interference regime for SU*. In this regime, the PU transmission power is weak and at a level lower than some threshold, such that the SU performance increases with decrease in detection probability. On the other hand, for the high interference regime in which the PU power is stronger, the SU cannot sacrifice the detection probability at all. This is because the interference caused by PU is so high that the SU performance decreases with decrease in detection probability for the high interference regime. Therefore, during this regime, the detection probability needs to be kept at the maximum level for the non-EDI region, such that it is greater than the required minimum threshold but small enough to avoid the EDI region. This change in the interference regimes for the SU performance is not revealed by the conventional model, and can only be observed under the mixture Gaussian model proposed in our study. Note that we also provide the values of the thresholds for the PU transmission power that determine the regime change for the SU performance, based on the numerical results. Furthermore, the discovery of such interesting trends is neither trivial nor straightforward. Our adopted optimization framework, explained in Sections 2.3.2 and 2.3.3, helps us identify these trends in SU performance. We finally demonstrate that even if the signal detector of SU performs suboptimally, SU can still achieve the optimal detector's performance level in the high interference regime for SU just by adjusting the sensing parameters accordingly. Our results provide a guideline for the design of a practical cognitive radio system, based on the transmission-power levels of the users, to achieve the maximum secondary-user's performance by keeping the primary-user's performance-degradation within a given tolerable range.

Since we are interested in revealing the effect of the mixture-Gaussian interference on the degrees-of-freedom and interference trade-off, we thus keep our model simple and do not incorporate fading and exact information of receivers' locations. However, these features are required for a more precise modeling of interference that is left for future exploration. Even in the absence of these features, the computational complexity of our problem is very high, due to the incoherence assumption between SU receiver and PU transmitter, which results in a mixture Gaussian in-

terference channel and non-linear interdependence of quite a few parameters in our system model, as explained in Section 2.3.2. Also recall that we only consider a single point-to-point communication link for both PU and SU in our system model. Even this simple model poses analytical challenges and reveals interesting results, as elaborated in this study. However, this basic model can be used as a building block for a more sophisticated model of cognitive radio networks, which may consist of large number of nodes and communication links among them. Such complicated model considerations are left for future work.

In Section 2.6, we also discuss how the analysis will be affected if we assume that the PU receiver (PU-RX) wrongly anticipates that the channel is Gaussian, and therefore, uses the *nearest neighbor decoding* rule, which is the optimal decoding rule for the Gaussian channel. Moreover, in Section 2.7, we discuss the analysis when the SU transmitter (SU-TX) uses a suboptimal detector, i.e., the envelope detector, to sense the PU signal. In the same section we also compare the results for the envelope detector to those for the optimal detector, i.e., the energy detector.

The rest of the chapter is organized as follows. We begin with presenting the system model in Section 3.2. Later, we formulate the optimization problem in Section 4.3. In Section 2.3.2 we present the optimization procedure for the formulated problem. We then compare our model with the conventional one in Section 2.4. Section 2.5 discusses the numerical results thoroughly. In Section 2.6 we discuss the scenario when nearest neighbor decoding scheme is employed by PU-RX. In Section 2.7 we present a comparison between the use of the optimal energy detector and the suboptimal envelope detector for SU-TX. Finally, we conclude in Section 4.7.

2.2 System Model

In Table 2.1 we present the details of the symbols used in this study. We consider a communication system, consisting of four nodes as shown in Figure 2.1. In our system model, SU-TX and SU-RX represent the secondary user transmitter and receiver, respectively, whereas, PU-TX and PU-RX represent the primary user transmitter and receiver respectively. In this study we consider periodic sensing that enables SU-TX to remain aware of the channel status. This is achieved by

TABLE 2.1: Table of Symbols

Symbol	Definition
N	Total number of channel uses in one frame
N_{s}	Number of channel uses in the sensing period of SU frame
N_{c}	Number of channel uses in the communication period of SU frame
τ	Portion of the total frame-time that is used by SU-TX for sensing
λ	SU-TX sensing threshold
π_0	Pr (PU-TX is off for the time of entire frame)
π_1	Pr (PU-TX is on for the time of entire frame)
ψ_1	Pr (SU-TX is on for the communication period of SU frame)
X_{PU}	Transmitted symbol during a single channel use by PU-TX
X_{SU}	Transmitted symbol during a single channel use by SU-TX
W_{PU}	Noise component in the received symbol for PU-RX during a single channel use
W_{SU}	Noise component in the received symbol for SU-RX during a single channel use
Y_{PU}	Received symbol for PU-RX during a single channel use
Y_{SU}	Received symbol for SU-RX during a single channel use
α	Coupling coefficient for PU-RX
β	Coupling coefficient for SU-RX
P_α	Pr ($\alpha = 1$)
P_β	Pr ($\beta = 1$)
$\mathcal{P}_{\mathrm{PU}}$	PU transmission power level
$\hat{\mathcal{P}}_{\mathrm{PU}}$	$\mathcal{P}_{\mathrm{PU}}$ threshold
$\mathcal{P}_{\mathrm{SU}}$	SU transmission power level
P_{F}	False alarm probability
P_{D}	Detection probability
$\tilde{P}_{\mathrm{D}}$	A given fixed value of P_{D}
$P_{\mathrm{D_T}}$	The constraint threshold for P_{D}
$P_{\mathrm{D_{T2}}}$	Value of P_{D} at which EDI region starts
H	Actual channel-state variable
$\hat{H}$	Decision channel-state variable
$I_{\mathrm{SU-N}}$	Mutual information for SU system for the entire frame
$\hat{I}_{\mathrm{SU-N}}$	SU information rate for a single channel use
$\tilde{I}_{\mathrm{SU-N}}$	$= \max_{\tau,\lambda} I_{\mathrm{SU-N}},\ \text{s.t. } P_{\mathrm{D}} = \tilde{P}_{\mathrm{D}}$
$\tilde{\hat{I}}_{\mathrm{SU-N}}$	Value of $\hat{I}_{\mathrm{SU-N}}$, when the SU information rate is $\tilde{I}_{\mathrm{SU-N}}$
ν_1	Upper bound on $I_{\mathrm{SU-N}}$ for perfect detection
$I_{\mathrm{PU-N}}$	Mutual information for PU system for the entire frame
$\tilde{I}_{\mathrm{PU-N}}$	$= I_{\mathrm{PU-N}},\ \text{s.t. } I_{\mathrm{SU-N}} = \tilde{I}_{\mathrm{SU-N}}$
$\hat{I}_{\mathrm{PU-Ns}}$	PU information rate for a single channel use during the sensing period of SU
$\hat{I}_{\mathrm{PU-Nc}}$	PU information rate for a single channel use during the communication period of SU
Δ_{PU}	PU performance degradation factor
Δ_{T}	Maximum allowed value of Δ_{PU}
$\tilde{\Delta}_0$	Value of Δ_{PU} achieved at optimal solution

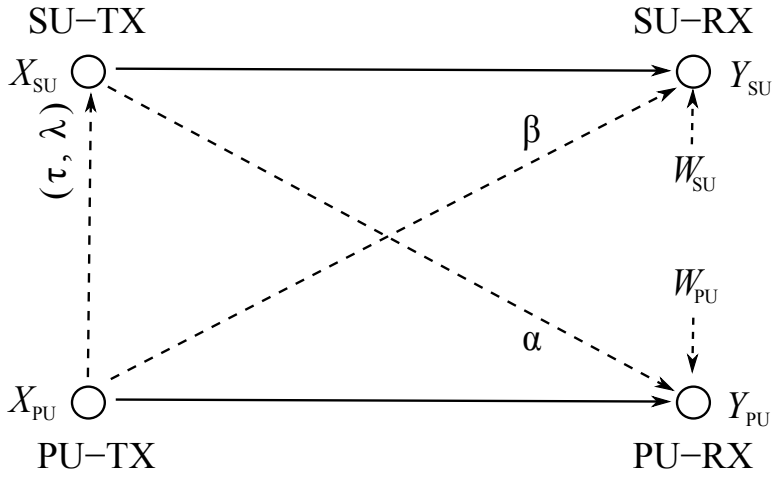


FIGURE 2.1: Cognitive radio system model.

using a frame structure [13, 25–27]. In this structure, each frame consists of a sensing period and a transmission period, while the overall length of the frame is fixed. At the end of each sensing period, the secondary transmission starts when the PU channel is considered as idle by SU-TX. Otherwise, SU-TX will wait until the next frame to sense the licensed channel again before any secondary usage. Thus we assume encoding of SU system is based on this frame structure such that each frame consists of, say, N channel uses and the information rate is achieved by encoding for infinitely large number of such frames.

We define a parameter τ as the portion of the total frame-time that is used by SU-TX for sensing the occupation of the channel by the PU. For SU-TX, $N_{\mathrm{s}} = \tau N$ represents the number of channel-uses in a frame out of a total of N that are used for sensing and is therefore called the *sensing period*. While $N_{\mathrm{c}} = N - N_{\mathrm{s}} = (1-\tau)N$ represents the number of channel-uses in a frame used for communications whenever there is a transmission opportunity for SU-TX and is thus called the *communication period* [13, 25–27].

The encoding of PU system is also based on a frame structure with N channel uses and the information rate is thus achievable for infinitely large number of transmissions of such frames. We assume that PU-TX communicates data to PU-RX in an on-and-off manner in the form of frames over the given frequency channel. This on-and-off operation of PU-TX is modeled such that for

every frame of N channel uses PU-TX is either off for the time of entire frame with probability π_0 or it is on for the time of entire frame with probability $\pi_1 = 1 - \pi_0$. It is assumed that there is a precedent learning or estimation phase where long-time scale measurements are taken to estimate π_1. Such learning or estimation problem is beyond the scope of this study, which will be further investigated and presented in our future works.

Let X_{PU} and X_{SU} denote the independent transmitted symbols during a single channel use by PU-TX and SU-TX respectively while they access the channel. Similarly, Y_{PU} and Y_{SU} denote the corresponding received symbols for PU-RX and SU-RX respectively. All communication channels in the system are assumed to be additive white Gaussian noise channels. Let W_{PU} and W_{SU} denote the independent noise components in the received symbols for PU-RX and SU-RX respectively. Both these noise components are Gaussian random variables having zero mean and unit variance. Also the transmitted signal and the noise, and therefore the received signal, over any given channel-use are independent but distributed identically to those over all the other channel-uses.

We assume that PU-TX and PU-RX have no knowledge about the existence of SU-TX and SU-RX in the system. Therefore, perceiving a simple additive Gaussian noise channel, PU-TX transmits a Gaussian distributed symbol X_{PU} over each single channel use thereby attempting to gain a maximum capacity. Note, in this work we are not aiming at finding the optimal probability distributions for X_{PU} and X_{SU} that can achieve the channel capacity in the presence of interference. Therefore, we assume both as Gaussian random variables. Thus, the transmitters are assumed to possess a Gaussian code-book. The transmitted symbols are real Gaussian random variables each of which has a zero mean and a variance that is equal to the average power constraint of the transmitter. We use $\mathcal{P}_{\mathrm{PU}}$ and $\mathcal{P}_{\mathrm{SU}}$ to denote the average power constraints for PU-TX and SU-TX respectively.

We adopt a channel sensing model for SU-TX that is based on the energy detection scheme. The false alarm probability P_{F} and the detection probability P_{D} for our energy detector based

sensing model are given by, [28], [29], [27], $P_F = \frac{\Gamma(N_S/2,\lambda/2)}{\Gamma(N_S/2)}$ and $P_D = \frac{\Gamma\left(N_S/2,\frac{\lambda}{2(1+\mathcal{P}_{\rm PU})}\right)}{\Gamma(N_S/2)}$. Here λ is the sensing threshold and, $\Gamma(.)$ and $\Gamma(.,.)$ are respectively the *complete* and *upper-incomplete* Gamma functions [30].

The received signals at PU-RX and SU-RX are given by $Y_{\rm PU} = X_{\rm PU} + W_{\rm PU} + \alpha X_{\rm SU}$ and $Y_{\rm SU} = X_{\rm SU} + W_{\rm SU} + \beta X_{\rm PU}$ respectively. Here, $X_{\rm PU}$ and $X_{\rm SU}$ are Gaussian distributed with probability density functions (PDFs) $\mathcal{N}(0, \mathcal{P}_{\rm PU})$ and $\mathcal{N}(0, \mathcal{P}_{\rm SU})$ respectively, while $W_{\rm PU}$ and $W_{\rm SU}$ both are Gaussian distributed with PDF $\mathcal{N}(0, 1)$. The random variables, α and β, are called the *coupling coefficients*. They appear because of the mutual interference between PU and SU users. We consider this interference in our system model because we assume that the SU-RX does not know the state of the channel access by PU-TX. Parameters α and β are Bernoulli random variables having a value 0 or 1. These coefficients can be described by the probabilities, $P_\alpha = \Pr(\alpha = 1)$ and $P_\beta = \Pr(\beta = 1)$, respectively, which are given by $P_\alpha = \Pr(\hat{H} = 0|H = 1) = 1 - P_{\rm D}$ and $P_\beta = \Pr(H = 1|\hat{H} = 0) = \frac{\pi_1(1-P_D)}{\pi_0(1-P_F)+\pi_1(1-P_D)}$. Here $\hat{H}$ and H denote the decision and actual channel state variables, respectively, with 0 representing an idle primary channel and 1 for a busy primary channel.

Using Bayes rule, the PDF of $Y_{\rm PU}$ is

$$f_{\rm Y_{PU}}(x) = (1 - P_\alpha) \times \mathcal{N}(x; 0, 1 + \mathcal{P}_{\rm PU}) + P_\alpha \times \mathcal{N}(x; 0, 1 + \mathcal{P}_{\rm PU} + \mathcal{P}_{\rm SU}), \tag{2.1}$$

and that of $Y_{\rm SU}$ is

$$f_{\rm Y_{SU}}(x) = (1 - P_\beta) \times \mathcal{N}(x; 0, 1 + \mathcal{P}_{\rm SU}) + P_\beta \times \mathcal{N}(x; 0, 1 + \mathcal{P}_{\rm PU} + \mathcal{P}_{\rm SU}). \tag{2.2}$$

It is evident from (2.1) and (2.2) that, because of the mutual interference between PU and SU the unwanted disturbance component in the received symbols have mixture Gaussian PDFs instead of pure Gaussian. Therefore we cannot use the well known result for the Gaussian channel capacity, $\log(1 + \rm SNR)$, as used for the conventional model where the unwanted component is simply Gaussian [13]. We therefore need to re-evaluate the information rates that are achievable over the resulting mixture Gaussian channel, in order to analyze the system performance.

For both SU-RX and PU-RX, the decoding schemes are assumed to be optimal, i.e., the Maximum Likelihood (ML) decoding. This assumption of using optimal decoding will ensure that the mutual information rates in our model are achievable and therefore considered as valid metrics for system throughput. But, it should be noted that ML decoding requires knowledge of channel statistics at the receiver and is complicated to implement for a non-Gaussian channel, as is our case. Later in Section 2.6, we comment on the effect of adopting a robust decoding scheme for PU-RX namely the *nearest neighbor decoding* [31].

2.3 Problem Formulation and Solution Procedure

2.3.1 Performance Metrics:

The performance metric for the SU system is the mutual information which quantifies the achievable communication rate per transmission frame, between SU-TX and SU-RX, when SU-RX uses the optimum maximum likelihood decoder, in the presence of an uncertainty of the PU channel state averaged over the long run, i.e., over infinitely large number of transmissions. Note that the PU channel state in our study represents whether the PU-TX is on or off. We assume that SU-RX does not have the knowledge of this PU channel state which causes interference in the SU system. Such a situation with a lack of knowledge on the channel quality status is called *incoherent* receiver operation [10–12]. Similarly, the performance metric for the PU system is the mutual information which quantifies the achievable communication rate per transmission frame, between PU-TX and PU-RX, in the presence of an uncertainty of the SU channel state averaged over the long run. Recall that the encoding of the SU and PU systems is based on a frame structure with N channel uses. The channel over each frame varies between good and bad, where good means no interference from another transmitter, and bad the opposite. The achievable rate is attainable through the encoding over a large number of frames to average out such channel uncertainties without delay constraints. If there is a memory between good and bad conditions, we essentially deal with the Giblert-Elliott channels [11], and such memory is essentially caused by the memory between on and off of PU-TX. In this study, we assume that PU-TX is either off over the entire

frame with probability π_0, or it is on with probability $\pi_1 = 1 - \pi_0$, without memory. Therefore, the mutual information for the PU system over N channel uses of the entire frame is given by

$$I_{\mathrm{PU-N}} = \pi_1 \left[N_{\mathrm{s}}\, \hat{I}_{\mathrm{PU-Ns}} + N_{\mathrm{c}}\, \hat{I}_{\mathrm{PU-Nc}} \right], \tag{2.3}$$

where $\hat{I}_{\mathrm{PU-Ns}}$ is the mutual information for PU over one channel use during the sensing period of SU-TX and $\hat{I}_{\mathrm{PU-Nc}}$ is the mutual information for PU over one channel use during the communication period of SU-TX. During the sensing period, SU-TX is always silent and the PU channel is simply an *Additive Gaussian Noise* (AGN) channel therefore $\hat{I}_{\mathrm{PU-Ns}} = (1/2)\log(1 + \mathcal{P}_{\mathrm{PU}})$. On the other hand, during the communication period SU-TX may transmit if it detects an opportunity and thus the PU channel is an *Additive Mixture Gaussian* (AMG) channel. Therefore,

$$\hat{I}_{\mathrm{PU-Nc}} = h(Y_{\mathrm{PU}}) - h(Y_{\mathrm{PU}}|X_{\mathrm{PU}}) = h(Y_{\mathrm{PU}}) - h(Z_{\mathrm{PU}}), \tag{2.4}$$

where $h(X) = -\int f_{\mathrm{X}}(x) \log\left[f_{\mathrm{X}}(x)\right] dx$ is the differential entropy of random variable X having PDF $f_{\mathrm{X}}(x)$ and $Z_{\mathrm{PU}} = W_{\mathrm{PU}} + \alpha X_{\mathrm{SU}}$ is a mixture Gaussian random variable with PDF

$$f_{Z_{\mathrm{PU}}}(x) = (1 - P_\alpha) \times \mathcal{N}(x; 0, 1) + P_\alpha \times \mathcal{N}(x; 0, 1 + \mathcal{P}_{\mathrm{SU}}). \tag{2.5}$$

SU system only communicates during the communication period whenever SU-TX detects an opportunity. Therefore, the mutual information for the SU system over N channel uses of the entire frame is given by

$$I_{\mathrm{SU-N}} = \psi_1\, N_{\mathrm{c}}\, \hat{I}_{\mathrm{SU-Nc}}. \tag{2.6}$$

Here the product term $\psi_1 N_{\mathrm{c}}$ represents the degrees of freedom available to SU for communications, which quantify the average number of channel-uses per transmission frame used by SU for communications, in the long run. Note that $\psi_1 = \Pr(\mathrm{SU\ is\ on}) = (1 - P_{\mathrm{F}})\pi_0 + (1 - P_{\mathrm{D}})\pi_1$ and $\hat{I}_{\mathrm{SU-Nc}}$ is the mutual information for SU over one channel use during the communication period, given as follows

$$\hat{I}_{\mathrm{SU-Nc}} = h(Y_{\mathrm{SU}}) - h(Y_{\mathrm{SU}}|X_{\mathrm{SU}}) = h(Y_{\mathrm{SU}}) - h(Z_{\mathrm{SU}}), \tag{2.7}$$

where $Z_{\mathrm{SU}} = W_{\mathrm{SU}} + \beta X_{\mathrm{PU}}$ is a mixture Gaussian random variable with PDF

$$f_{\mathrm{Z_{SU}}}(x) = (1 - P_\beta) \times \mathcal{N}(x; 0, 1) + P_\beta \times \mathcal{N}(x; 0, 1 + \mathcal{P}_{\mathrm{PU}}). \tag{2.8}$$

2.3.2 Optimization Problem and Procedure for Solution:

A low detection probability requires small sensing time and thus corresponds to high degrees of freedom available to SU for communications. At the same time, it causes high interference for both SU and PU that results in a degradation in performance of both systems, and vice-versa. Therefore, variations in the sensing time and threshold affect the detection probability which incurs both loss and gain at the same time for SU, thereby indicating the need of a trade-off for SU performance. Our objective is to find the optimal sensing threshold λ and sensing time τ that will maximize the SU performance and at the same time keep the detection quality within a suitable range. Thus, our objective is to find the optimal τ and λ that will maximize the mutual information $I_{\mathrm{SU-N}}$ under the constraint that P_{D} will not drop below a given threshold $P_{\mathrm{D_T}}$, i.e.,

$$\max_{\tau,\lambda} I_{\mathrm{SU-N}}, \text{ s.t. } P_{\mathrm{D}} \geq P_{\mathrm{D_T}}. \tag{2.9}$$

Based on our discussion in Section 3.2 and 2.3.1, we can express the objective function in (4.31) in a more elaborative form as follows

$$\max_{\tau,\lambda} \ \psi_1\Big(P_{\mathrm{D}}(\tau,\lambda), P_{\mathrm{F}}(\tau,\lambda)\Big).(1-\tau)N. \int_{x\in\mathbb{R}} g\Big(x, P_{\mathrm{D}}(\tau,\lambda), P_{\mathrm{F}}(\tau,\lambda)\Big)dx, \ \text{ s.t. } P_{\mathrm{D}}(\tau,\lambda) \geq P_{\mathrm{D_T}}. \tag{2.10}$$

Note that $g(x, P_{\mathrm{D}}, P_{\mathrm{F}}) = f_{\mathrm{Y_{SU}}}(x) - f_{\mathrm{Z_{SU}}}(x)$, whereas the integral $\int g(x, P_{\mathrm{D}}, P_{\mathrm{F}})dx = \hat{I}_{\mathrm{SU-Nc}}$ is a difference of entropies of mixture Gaussian random variables, as shown in (2.7), which does not have a closed form solution and needs to be evaluated numerically [32]. Moreover, as explained in Section 3.2, ψ_1 in (2.10) and the distributions of the mixture Gaussian variables in our model are functions of false alarm and detection probabilities given by $P_F = \frac{\Gamma(N_S/2, \lambda/2)}{\Gamma(N_S/2)}$ and $P_D = \frac{\Gamma\left(N_S/2, \frac{\lambda}{2(1+\mathcal{P}_{\mathrm{PU}})}\right)}{\Gamma(N_S/2)}$. Recall from Section 3.2 that these probabilities are ratios of Gamma functions which are integrals without closed form solution [30]. Due to this analytical complexity

of problem (4.31), we cannot use the standard analytical methods for optimization. Therefore, in this study, we develop a numerical framework to solve the optimization problem (4.31), which is discussed in Section 2.5. The framework includes the numerical evaluation of entropy of mixture Gaussian distributions and the numerical solution of problem (4.31). Here, in this Section, we present our two step optimization procedure that is used in the numerical framework. This optimization procedure enables us to quantify the SU performance as a function of a single variable, instead of two variables τ and λ, as explained in the subsequent paragraph. This helps generate SU performance results that are easy to visualize and analyze, and also present useful insights, as discussed in Section 2.5.

In the first step of optimization, we fix P_{D} to some value say $\tilde{P}_{\mathrm{D}}$. For a given P_{D}, $I_{\mathrm{SU-N}}$ varies with P_{F}, as detailed in Sections 3.2 and 4.3. So, we maximize $I_{\mathrm{SU-N}}$ over all such τ and λ for which $P_{\mathrm{D}} = \tilde{P}_{\mathrm{D}}$. This first step results in the maximum values of $I_{\mathrm{SU-N}}$ for each $\tilde{P}_{\mathrm{D}}$, which we represent as $\tilde{I}_{\mathrm{SU-N}}$, given by

$$\tilde{I}_{\mathrm{SU-N}} = \max_{\tau,\lambda} I_{\mathrm{SU-N}}, \text{ s.t. } P_{\mathrm{D}} = \tilde{P}_{\mathrm{D}}. \tag{2.11}$$

$\tilde{I}_{\mathrm{SU-N}}$ is thus a function of $\tilde{P}_{\mathrm{D}}$. Note, that $\tilde{P}_{\mathrm{D}}$ is used only for expressing the mathematical formulation (2.11), and that $\tilde{P}_{\mathrm{D}}$ and P_{D} are interchangeable everywhere in our analysis. In the second step of our optimization procedure, we select the maximum $\tilde{I}_{\mathrm{SU-N}}$ such that $\tilde{P}_{\mathrm{D}} \geq P_{\mathrm{D_T}}$ as shown below in (2.12).

$$\max_{\tilde{P}_{\mathrm{D}}} \tilde{I}_{\mathrm{SU-N}}, \text{ s.t. } \tilde{P}_{\mathrm{D}} \geq P_{\mathrm{D_T}}. \tag{2.12}$$

This concludes the optimization procedure. Note that the two step optimization procedure enables us to quantify the SU performance in terms of $\tilde{I}_{\mathrm{SU-N}}$ which is a function of a single variable $\tilde{P}_{\mathrm{D}}$, instead of $I_{\mathrm{SU-N}}$ which is a function of two variables τ and λ and therefore not easy to analyze. The SU performance, in terms of $\tilde{I}_{\mathrm{SU-N}}$, is studied in detail in Section 2.5.

2.3.3 The Constraint Threshold $P_{\mathrm{D_T}}$:

We have already introduced $\tilde{I}_{\mathrm{SU-N}}$, a function of $\tilde{P}_{\mathrm{D}}$, as the performance metric for SU. It is the maximum value of $I_{\mathrm{SU-N}}$ over all such τ and λ for which $P_{\mathrm{D}} = \tilde{P}_{\mathrm{D}}$. For the same detection probability, the corresponding value of $I_{\mathrm{PU-N}}$ provides the measure of PU performance and is represented as $\tilde{I}_{\mathrm{PU-N}}$, defined as follows

$$\tilde{I}_{\mathrm{PU-N}} = I_{\mathrm{PU-N}},\ \text{s.t.}\ P_{\mathrm{D}} = \tilde{P}_{\mathrm{D}}. \tag{2.13}$$

Thus $\tilde{I}_{\mathrm{PU-N}}$, a function of $\tilde{P}_{\mathrm{D}}$, is the performance metric for PU. Also, for all such τ and λ for which $P_{\mathrm{D}} = \tilde{P}_{\mathrm{D}}$, $I_{\mathrm{PU-N}}$ is a constant as detailed in Section 3.2 and 4.3. Therefore we do not need to maximize it as we did for $\tilde{I}_{\mathrm{SU-N}}$.

A decrease in detection probability always deteriorates the PU performance due to increase in interference by SU. Thus, $\tilde{I}_{\mathrm{PU-N}}$ always decreases with decrease in $\tilde{P}_{\mathrm{D}}$. Therefore, for the optimization problem (2.12), at first, we select an appropriate detection probability threshold $P_{\mathrm{D_T}}$ such that the PU performance $\tilde{I}_{\mathrm{PU-N}}$ does not drop below a given threshold. For this, we define the *PU performance degradation* (PPD) metric, for a certain value of $\tilde{I}_{\mathrm{PU-N}}$, with respect to the value of $\tilde{I}_{\mathrm{PU-N}}$ at $\tilde{P}_{\mathrm{D}} = 1$. This metric is given as follows

$$\Delta_{\mathrm{PU}} = \frac{\tilde{I}_{\mathrm{PU-N}}(P_{\mathrm{D}} = 1) - \tilde{I}_{\mathrm{PU-N}}}{\tilde{I}_{\mathrm{PU-N}}(P_{\mathrm{D}} = 1)}. \tag{2.14}$$

Thus, for some given maximum value of PU performance degradation factor say Δ_{T}, we can find the minimum eligible value of $\tilde{I}_{\mathrm{PU-N}}$. We evaluate $P_{\mathrm{D_T}}$ as the value of $\tilde{P}_{\mathrm{D}}$ corresponding to this minimum eligible PU rate. Therefore, we can write

$$P_{\mathrm{D_T}} = \tilde{P}_{\mathrm{D}},\ \text{s.t.}\ \Delta_{\mathrm{PU}} = \Delta_{\mathrm{T}}. \tag{2.15}$$

In this way, we find an appropriate value for the constraint threshold $P_{\mathrm{D_T}}$ that ensures the PU performance degradation remains within a given tolerable range. Later, for all $\tilde{P}_{\mathrm{D}}$ values greater than or equal to $P_{\mathrm{D_T}}$, we maximize the SU performance $\tilde{I}_{\mathrm{SU-N}}$. As a result, we eventually reach

our goal of maximizing the SU performance for an acceptable degradation in PU performance, as summarized below.

$$\max_{\tilde{P}_{\mathrm{D}}} \tilde{I}_{\mathrm{SU-N}},\ \text{s.t. } \tilde{P}_{\mathrm{D}} \geq P_{\mathrm{D_T}},$$

$$\text{where, } \tilde{I}_{\mathrm{SU-N}} = \max_{\tau,\lambda} I_{\mathrm{SU-N}},\ \text{s.t. } P_{\mathrm{D}} = \tilde{P}_{\mathrm{D}},$$

$$\text{and, } P_{\mathrm{D_T}} = \tilde{P}_{\mathrm{D}},\ \text{s.t. } \Delta_{\mathrm{PU}} = \Delta_{\mathrm{T}}.$$

2.4 Comparison with the Conventional Model

To the best of the authors' knowledge the studies done so far on sensing-throughput trade-off have implicitly or explicitly made an important assumption that the SU-RX operates coherently with PU-TX. Under this assumption the received signal at SU-RX is given by

$$Y_{\mathrm{SU}} = \begin{cases} X_{\mathrm{SU}} + W_{\mathrm{SU}} & \text{if } \hat{H} = 0, H = 0. \\ X_{\mathrm{SU}} + W_{\mathrm{SU}} + X_{\mathrm{PU}}, & \text{if } \hat{H} = 0, H = 1. \end{cases} \tag{2.16}$$

While, the SU information rate for this model is given by

$$\begin{aligned} I_{\mathrm{SU-N}} &= N_c\ I(X_{\mathrm{SU}}; Y_{\mathrm{SU}} | \hat{H}, H) \\ &= (1-\tau)N \\ &\times [\Pr(\hat{H} = 0, H = 0)\ I(X_{\mathrm{SU}}; Y_{\mathrm{SU}} | \hat{H} = 0, H = 0) \\ &+ \Pr(\hat{H} = 0, H = 1)\ I(X_{\mathrm{SU}}; Y_{\mathrm{SU}} | \hat{H} = 0, H = 1)]. \end{aligned} \tag{2.17}$$

For this model the capacity can be achieved if X_{PU} and X_{SU} are Gaussian distributed with probability density functions (PDFs) $\mathcal{N}(0, \mathcal{P}_{\mathrm{PU}})$ and $\mathcal{N}(0, \mathcal{P}_{\mathrm{SU}})$ respectively, with W_{PU} and W_{SU} both are Gaussian distributed with PDF $\mathcal{N}(0, 1)$. Under these conditions, the achievable SU information becomes [13, 20–26],

$$I_{\mathrm{SU-N}} = (1-\tau)N \times [\pi_0\ (1 - P_{\mathrm{F}})\ \log(1 + \mathcal{P}_{\mathrm{SU}}) + \pi_1\ (1 - P_{\mathrm{D}})\ \log(1 + \frac{\mathcal{P}_{\mathrm{SU}}}{1 + \mathcal{P}_{\mathrm{PU}}})]. \tag{2.18}$$

In the conventional studies, optimal τ and λ are found to maximize $I_{\mathrm{SU-N}}$. Thus, the conventional optimization problem for the sensing-throughput trade-off is given by

$$\max_{\tau,\lambda} I_{\mathrm{SU-N}},\ \text{s.t. } P_{\mathrm{D}} \geq P_{\mathrm{D_T}}. \tag{2.19}$$

In the conventional model, the assumption that SU-RX operates coherently with PU-TX has major consequences on the sensing-throughput trade-off. Due to this assumption SU information rates for a single channel use, $I(X_{\mathrm{SU}}; Y_{\mathrm{SU}}|\hat{H}=0, H=0) = \log(1+\mathcal{P}_{\mathrm{SU}})$ and $I(X_{\mathrm{SU}}; Y_{\mathrm{SU}}|\hat{H}=0, H=1) = \log(1+\frac{\mathcal{P}_{\mathrm{SU}}}{1+\mathcal{P}_{\mathrm{PU}}})$, are independent of P_{D}. Now, since $P_{\mathrm{F}} \leq P_{\mathrm{D}}$ always holds, therefore according to (2.18) $I_{\mathrm{SU-N}}$ is a decreasing function of P_{D} for a fixed τ. Therefore, for the conventional model, $P_{\mathrm{D}} \geq P_{\mathrm{D_T}}$ constraint is equivalent to $P_{\mathrm{D}} = P_{\mathrm{D_T}}$. Hence, conventional model only requires maximizing $I_{\mathrm{SU-N}}$ with respect to τ for an arbitrarily high value of $P_{\mathrm{D_T}}$.

As a contrast, in our model, we assume that SU-RX is not coherent to PU-TX. Consequently, we have a mixture Gaussian channel, with a more complicated coupling relationship between the primary and the secondary users. Such coupling relationship is reflected not only by the presence or absence of an interference component in the received signals, which already showed up in conventional models, but also in a more complicated computation of the mutual information under our proposed mixture model, as explained in Section 2.3.2. This has two implications. Firstly, SU information rate $\hat{I}_{\mathrm{SU-N}}$ for a single channel use depends on P_{D} for a fixed τ so that $P_{\mathrm{D}} \geq P_{\mathrm{D_T}}$ constraint is not equivalent to $P_{\mathrm{D}} = P_{\mathrm{D_T}}$ for our optimization problem. Thus we first maximize $I_{\mathrm{SU-N}}$ with respect to τ for any given $P_{\mathrm{D}} = \tilde{P}_{\mathrm{D}}$ to seek a relationship over all. Secondly, the effect of P_{D} on SU performance depends on the interference caused by PU. The level of this interference depends on $\mathcal{P}_{\mathrm{PU}}$ for a fixed $\mathcal{P}_{\mathrm{SU}}$. Due to this, we have identified interference regimes for the SU performance based on $\mathcal{P}_{\mathrm{PU}}$ level, as discussed in Section 2.5.1. Also, in our analysis, instead of assigning an arbitrary value to the optimization problem constraint $P_{\mathrm{D_T}}$ we assign a value based on the tolerable range of PU performance degradation as discussed in Section 2.3.2.

2.5 Results and Discussion

This section discusses the relationship between the degrees-of-freedom (available to SU for communications) and the interference experienced by SU. This interference is caused by the primary user signal due to imperfection in detection by SU and it affects the SU channel-information-rate. Finally at the end of this section, we present and discuss the solution to the optimization problem under consideration.

As discussed in Sections 3.2 and 4.3, the resulting interference channel for our system model is mixture Gaussian and in order to evaluate the information rates required for the analysis we need to compute the differential entropy of mixture Gaussian random variables. However, the differential entropy of a mixture Gaussian random variable does not have a closed form solution. Moreover, the optimization problem formulated in Section 4.3 involves quite a few parameters that are also non-linearly interdependent. Therefore, we have solved the optimization problem numerically. For the numerical evaluation of differential entropy of a mixture Gaussian random variable, we have implemented an algorithm based on the methodology proposed in [32]. In this section, we present the numerical results for $\pi_1 = 0.7$, $N = 100$ and $\mathcal{P}_{\mathrm{SU}} = 10$. Also, in Section 2.5.3 we present the optimal solution for two values of $\mathcal{P}_{\mathrm{SU}}$, i.e., 0.1 and 10.

2.5.1 Objective function ($\tilde{I}_{SU-N}$) versus $\tilde{P}_{\mathrm{D}}$

The SU performance metric is $\tilde{I}_{\mathrm{SU-N}}$ that is the maximum $I_{\mathrm{SU-N}}$ at $P_{\mathrm{D}} = \tilde{P}_{\mathrm{D}}$. $\tilde{I}_{\mathrm{SU-N}}$ is thus a function of $\tilde{P}_{\mathrm{D}}$ as discussed in Section 2.3.2. Here we discuss this SU performance metric as a function of $\tilde{P}_{\mathrm{D}}$ in detail in order to elaborate and explain the trade-off between the degrees of freedom available to SU for communications and the interference experienced by SU due to PU.

Degrees of freedom available to SU: For the SU system, the product $N_{\mathrm{c}}\psi_1$ that appears in (2.6) is the measure of the *degrees of freedom* (DOF) for communications. It quantifies the average number of channel-uses per frame used by SU for communications, in the long run, i.e., for infinitely large number of frames. Recall that we introduced the notions of 'channel-use' and 'frame' in Section 3.2, where we described a synchronized frame structure for transmission of

information in the system. The values of this DOF measure such that $P_{\mathrm{D}} = \tilde{P}_{\mathrm{D}}$ and the SU information rate is $\tilde{I}_{\mathrm{SU-N}}$, are plotted against $\tilde{P}_{\mathrm{D}}$ in Figure 2.2. In this figure we present these numerical plots for different values of PU transmission power, $\mathcal{P}_{\mathrm{PU}}$. The figure shows that DOF

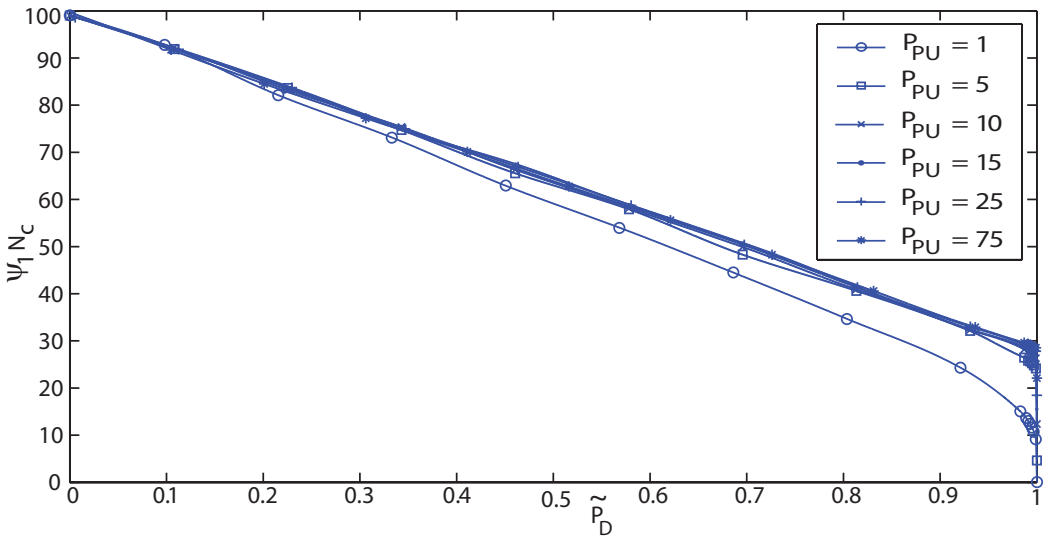


FIGURE 2.2: DOF available for SU for communications.

available for SU decreases as the detection probability increases for all values of $\mathcal{P}_{\mathrm{PU}}$. This is because, increase in detection probability requires an increase in sensing time τ that results in a decrease in the communication period available for SU that is $N_{\mathrm{c}} = (1-\tau)N$. At the same time, the increase in τ also causes an increase in false alarm probability P_{F}, which in turn causes an increase in the detection probability. Altogether, this decreases the probability that SU is ON, i.e., $\psi_1 = \Pr(\text{SU is on}) = (1-P_{\mathrm{F}})\pi_0 + (1-P_{\mathrm{D}})\pi_1$. Hence, we conclude, that an increase in detection probability corresponds to a decrease in both N_{c} and ψ_1 thereby causing a decrease in DOF available for SU.

Interference caused by PU: In (2.6), $\hat{I}_{\mathrm{SU-N}}$ is the mutual information of the SU system for a single channel use. It is the indicator for the interference caused by the PU system. The greater the interference, the lower will be the value of $\hat{I}_{\mathrm{SU-N}}$. We represent the values of $\hat{I}_{\mathrm{SU-N}}$, when $P_{\mathrm{D}} = \tilde{P}_{\mathrm{D}}$ and the SU information rate is $\tilde{I}_{\mathrm{SU-N}}$, as $\tilde{\hat{I}}_{\mathrm{SU-N}}$, i.e., with both tilde and a hat. $\tilde{\hat{I}}_{\mathrm{SU-N}}$ is plotted against $\tilde{P}_{\mathrm{D}}$ in Figure 2.3. In this figure we present these numerical plots for different

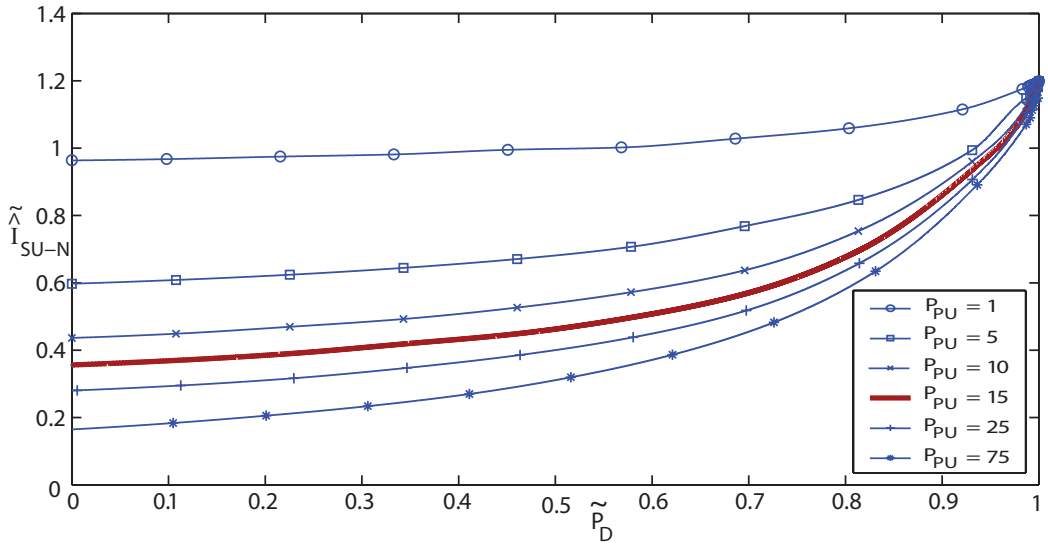


FIGURE 2.3: SU information for single channel-use vs $\tilde{P}_{\mathrm{D}}$.

values of PU transmission power, $\mathcal{P}_{\mathrm{PU}}$. The level of interference in the SU system caused by the PU is determined by two factors, namely, the level of PU-TX transmission power $\mathcal{P}_{\mathrm{PU}}$ and the SU-TX detection quality that is characterized by the value of $\tilde{P}_{\mathrm{D}}$. Higher values of $\mathcal{P}_{\mathrm{PU}}$ and lower values of $\tilde{P}_{\mathrm{D}}$ result in high interference levels thereby resulting in lower values of $\tilde{\hat{I}}_{\mathrm{SU-N}}$ in the SU system, and vice-versa. Thus, for a fixed $\mathcal{P}_{\mathrm{PU}}$, $\tilde{\hat{I}}_{\mathrm{SU-N}}$ increases with increase in $\tilde{P}_{\mathrm{D}}$, as shown in Figure 2.3. Also for a fixed $\tilde{P}_{\mathrm{D}} \neq 1$, $\tilde{\hat{I}}_{\mathrm{SU-N}}$ is lower for higher values of $\mathcal{P}_{\mathrm{PU}}$, as shown in Figure 2.3. But, for $\tilde{P}_{\mathrm{D}} = 1$, PU and SU do not interfere at all and both PU and SU channels are simply Gaussian. In that case, $\tilde{\hat{I}}_{\mathrm{SU-N}} = (1/2)\log(1+\mathcal{P}_{\mathrm{SU}})$ and is independent of $\mathcal{P}_{\mathrm{PU}}$. Therefore, all the curves meet at a single point at $\tilde{P}_{\mathrm{D}} = 1$.

Degrees of freedom versus interference trade-off for SU: For SU system, on one hand the increase in detection probability causes a decrease in DOF available to SU for communications and on the other hand it causes an increase in the information rate for a single channel use. The overall resulting information rate for SU, i.e., $\tilde{I}_{\mathrm{SU-N}}$, is the product of DOF available and the information rate for a single channel use $\tilde{\hat{I}}_{\mathrm{SU-N}}$. Hence, there exists a trade-off for the SU system between the available DOF and the interference level. This trade-off determines the overall

trends in SU performance. The SU performance curves, for different values of $\mathcal{P}_{\rm PU}$, are shown in Figure 2.4. These curves are the numerical plots of $\tilde{I}_{SU-N}$ versus $\tilde{P}_{\rm D}$.

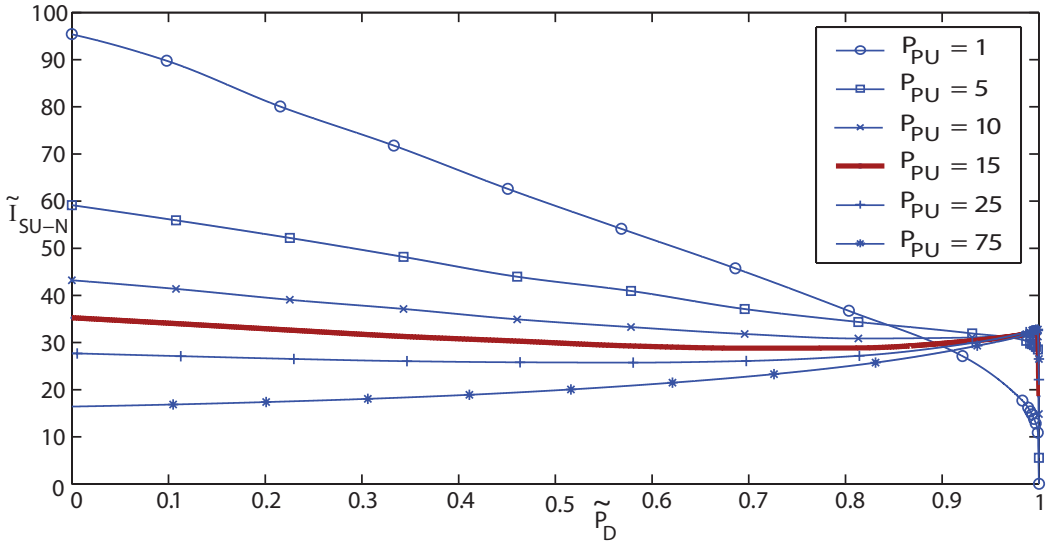


FIGURE 2.4: SU performance curves.

Energy detector's inefficiency region: Near $\tilde{P}_{\rm D} = 1$, a region of sharp decline in $\tilde{I}_{\rm SU-N}$ exists, which we call as the *energy detector's inefficiency* (EDI) region. This sharp decline in SU performance is due to the energy detection scheme's behavior depending on the gamma function. Near $\tilde{P}_{\rm D} = 1$, a very high increase in τ is required for a very small increase in $\tilde{P}_{\rm D}$. The communication period $N_{\rm c}$ thus decreases quite significantly with τ, thereby causing a sharp decrease in the degrees of freedom for communications as shown in Figure 2.2. During this EDI region, $\tilde{P}_{\rm D}$ is very high, i.e., nearly 1. Therefore, in this high-detection-probability region, the interference caused by the PU system is very negligible and does not affect $\tilde{\tilde{I}}_{\rm SU-N}$ much, no matter what the PU transmission power level is, as depicted in Figure 2.3. Hence, in this region the SU channel-rate $\tilde{\tilde{I}}_{\rm SU-N}$ is almost constant but the steep decline in DOF causes a similar decline in SU performance, as suggested by (2.6). For all selected values of $\mathcal{P}_{\rm PU}$, the EDI region starts almost at the same value of $\tilde{P}_{\rm D}$ defined as $P_{\rm D_{T2}}$. Numerical results, as presented in Figures 2.4 and 2.5, show that if $\tilde{P}_{\rm D}$ exceeds $P_{\rm D_{T2}}$ and enters the EDI region, then it will not improve the PU performance much, but deteriorate the SU performance quite significantly, even when we are increasing the detection probability.

This is what we mean by the word 'inefficiency' here. Thus, $\tilde{I}_{\mathrm{SU-N}}(P_{\mathrm{D_{T2}}}) >> \tilde{I}_{\mathrm{SU-N}}(1)$ while $\tilde{I}_{\mathrm{PU-N}}(P_{\mathrm{D_{T2}}}) \approx \tilde{I}_{\mathrm{PU-N}}(1)$. Thus, it is better to avoid the inefficient EDI region by sacrificing some $\tilde{P}_{\mathrm{D}}$ in order to have a better SU performance at almost no cost of PU performance. Hence, the appropriate value of $\tilde{P}_{\mathrm{D}}$ is upper bounded by $P_{\mathrm{D_{T2}}}$ instead of being exactly equal to 1. Our numerical algorithm can find the value of $P_{\mathrm{D_{T2}}}$ for some higher levels of PU transmission power as discussed in remark 1 in Section 2.5.3.

Upper bound of SU performance for perfect detection: Consider the case of perfect detection, i.e., $P_{\mathrm{D}} = 1$. For this case, PU and SU systems do not interfere with each other and both PU and SU channels are simply Gaussian. Therefore, at $P_{\mathrm{D}} = 1$, the SU information rate for a single channel use is $\hat{I}_{\mathrm{SU-N}}\ (P_{\mathrm{D}} = 1) = (1/2) \times \log(1 + \mathcal{P}_{\mathrm{SU}})$. Also, in the case of perfect detection, SU remains ON only when PU is OFF and there is no false alarm, i.e., the noise power level remains below the sensing threshold level λ. In this scenario, probability that SU is ON is thus given by $\psi_1(P_{\mathrm{D}} = 1) = [1 - \pi_1] \times [1 - P_{\mathrm{F}}]$. Now consider a limiting case of infinitely high PU transmission power, to achieve perfect detection. In this case, SU-TX requires infinitesimally small sensing time, i.e., $\tau \approx 0$ to achieve perfect detection. Also due to very high PU transmission power, the false alarm probability can be kept infinitesimally small at the same time, by selecting a very large sensing threshold and thus $P_{\mathrm{F}} \approx 0$. For such a scenario, the upper bound on $I_{\mathrm{SU-N}}$ for perfect detection can be achieved and is evaluated as $\nu_1 = \frac{1}{2}(1-\pi_1)N \log(1+\mathcal{P}_{\mathrm{SU}})$, using (2.6).

PU transmission power threshold: As shown in Figure 2.4, at $\tilde{P}_{\mathrm{D}} = 1$, $\tilde{I}_{\mathrm{SU-N}}$ increases with increase in PU transmission power $\mathcal{P}_{\mathrm{PU}}$ but remains below ν_1 since it is the upper bound on the SU information rate at $P_{\mathrm{D}} = 1$. Now, consider the case of $\tilde{P}_{\mathrm{D}} = 0$. On one hand, for higher values of PU transmission powers, $\tilde{I}_{\mathrm{SU-N}}$ is less than ν_1. On the other hand, for certain lower values of PU transmission powers, $\tilde{I}_{\mathrm{SU-N}}(\tilde{P}_{\mathrm{D}} = 0)$ is greater than ν_1. This motivates us to categorize SU performance curves on the basis of SU performance in case of no detection, i.e., when $P_{\mathrm{D}} = 0$. We therefore define a threshold value of $\mathcal{P}_{\mathrm{PU}}$ represented as $\hat{\mathcal{P}}_{\mathrm{PU}}$ such that $\tilde{I}_{\mathrm{SU-N}}(\tilde{P}_{\mathrm{D}} = 0) = \nu_1$. We called the region of $\tilde{P}_{\mathrm{D}} \geq P_{\mathrm{D_{T2}}}$ as the *EDI Region* for SU. Thus the region of $\tilde{P}_{\mathrm{D}} \leq P_{\mathrm{D_{T2}}}$ is

the *Non-EDI Region* for SU and consists of two regimes, namely the *High and Low Interference Regimes* for the SU. We define these regimes on the basis of PU transmission power level, as compared to the threshold $\hat{\mathcal{P}}_{\mathrm{PU}}$ that we have just introduced. These regimes are discussed in the subsequent discussion. We numerically estimate the value of $\hat{\mathcal{P}}_{\mathrm{PU}}$. For the system parameters that we have selected for our numerical study, $\hat{\mathcal{P}}_{\mathrm{PU}} \approx 15$. We call the corresponding SU performance curve as the *Neutral* $\tilde{I}_{\mathrm{SU-N}}$ *Curve*. It shows negligible variations with respect to $\tilde{P}_{\mathrm{D}}$, indicating an underlying balance for the DOF and interference trade-off for SU, for the PU transmission-power level of $\hat{\mathcal{P}}_{\mathrm{PU}}$.

Interference regimes for SU (non-EDI region): The non-EDI region with a weak primary power, i.e., $\mathcal{P}_{\mathrm{PU}} < \hat{\mathcal{P}}_{\mathrm{PU}}$, is called the *Low Interference Regime*. The $\tilde{I}_{\mathrm{SU-N}}$ versus $\tilde{P}_{\mathrm{D}}$ curves, above the Neutral $\tilde{I}_{\mathrm{SU-N}}$ Curve, lie in the Low Interference Regime for SU. There are two different trends of SU performance in this regime. The first one is for the values of $\mathcal{P}_{\mathrm{PU}}$ close to $\hat{\mathcal{P}}_{\mathrm{PU}}$ in which $\tilde{I}_{\mathrm{SU-N}}$ first decreases and then increases with increase in $\tilde{P}_{\mathrm{D}}$, for the non-EDI region as shown in Figure 2.4. This trend constitutes the *Stronger Low-Interference Regime*. The second trend is for the lower values of $\mathcal{P}_{\mathrm{PU}}$ for which $\tilde{I}_{\mathrm{SU-N}}$ decreases with increase in $\tilde{P}_{\mathrm{D}}$ for the non-EDI region as shown in Figure 2.4. This is because the effect of DOF dominates that of the interference in this regime of SU operation. As shown in Figure 2.2, for all values of $\mathcal{P}_{\mathrm{PU}}$, DOF decreases with increase in $\tilde{P}_{\mathrm{D}}$. Therefore in this Non-EDI region, $\tilde{I}_{\mathrm{SU-N}}$ decreases with increase in $\tilde{P}_{\mathrm{D}}$. This trend constitutes the *Weaker Low-Interference Regime*. The non-EDI region with a stronger primary power, i.e., $\mathcal{P}_{\mathrm{PU}} > \hat{\mathcal{P}}_{\mathrm{PU}}$, is called the *High Interference Regime*. The effect of interference dominates that of the DOF in this regime of SU operation. As shown in Figure 2.3, for all values of $\mathcal{P}_{\mathrm{PU}}$, $\tilde{\hat{I}}_{\mathrm{SU-N}}$ increases with increase in $\tilde{P}_{\mathrm{D}}$, because of the decrease in interference within the SU system due to PU. Therefore in this Non-EDI region, $\tilde{I}_{\mathrm{SU-N}}$ increases with increase in $\tilde{P}_{\mathrm{D}}$. The $\tilde{I}_{\mathrm{SU-N}}$ versus $\tilde{P}_{\mathrm{D}}$ curves, below the Neutral $\tilde{I}_{\mathrm{SU-N}}$ Curve, lie in the High Interference Regime for SU. Note that such interesting trends in SU performance, based on the interference

caused by PU, are not revealed by the conventional model and can only be observed under the mixture Gaussian model proposed in our study.

2.5.2 $\tilde{I}_{PU-N}$ versus $\tilde{P}_{\mathrm{D}}$

The $\tilde{I}_{PU-N}$ versus $\tilde{P}_{\mathrm{D}}$ curves are the PU performance curves. Figure 2.5 shows that $\tilde{I}_{PU-N}$ increases with $\tilde{P}_{\mathrm{D}}$ for all values of $\mathcal{P}_{\mathrm{PU}}$. The trends in these curves are because of the interference in the PU system due to SU. The effect of this interference in PU system depends on the level of $\mathcal{P}_{\mathrm{PU}}$ and the value of $\tilde{P}_{\mathrm{D}}$. For a given $\mathcal{P}_{\mathrm{PU}}$, an increase in $\tilde{P}_{\mathrm{D}}$ causes a decrease in interference from SU and thus PU performance increases. On the other hand, for a given $\tilde{P}_{\mathrm{D}}$, increase in $\mathcal{P}_{\mathrm{PU}}$ increases the PU information rate thereby increasing the PU performance.

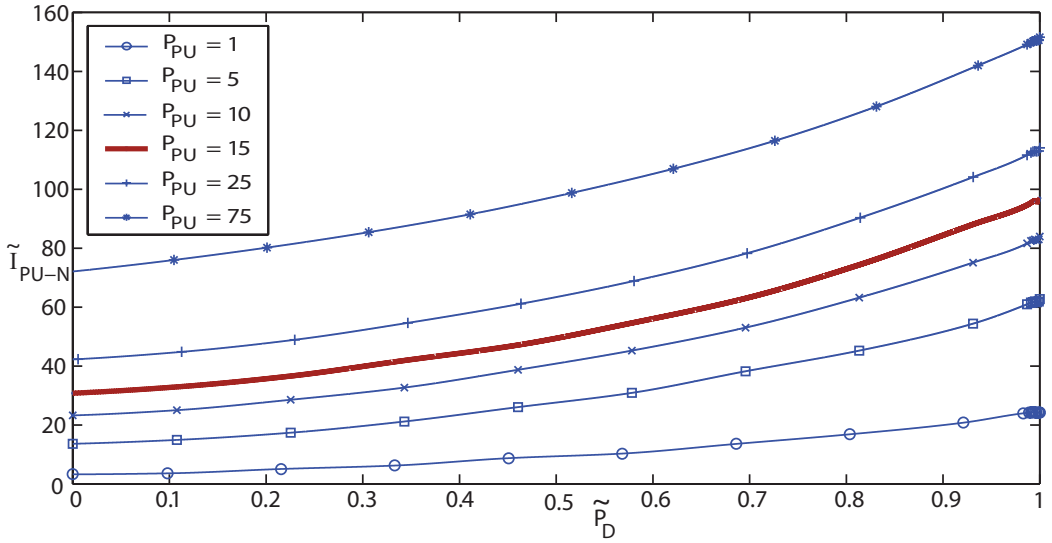


FIGURE 2.5: PU performance curves.

2.5.3 Optimal Solution

Based on the optimization procedure suggested in Section 2.3.2 we numerically solve problem (4.31). For quite a few selected values of $\mathcal{P}_{\mathrm{PU}}$ the results for $\Delta_{\mathrm{T}} = 0.04$ are given in Table 2.2 and 2.3. Here in these tables, $\tilde{P}_{\mathrm{D0}}$ and $\tilde{\Delta}_0$ are the values of detection probability and Δ_{PU}, respectively, for the corresponding optimal solution. Following are some remarks related to the numerical solutions presented in Table 2.2 and 2.3.

TABLE 2.2: Optimized Solutions for $\Delta_{\mathrm{T}} = 0.04$, $\mathcal{P}_{\mathrm{SU}} = 10$, Energy Detector

$\mathcal{P}_{\mathrm{PU}}$	max I_{SU}	$\tilde{\tau}_0$	$\tilde{\lambda}_0$	$\tilde{P}_{\mathrm{D0}}$	$\tilde{P}_{\mathrm{F0}}$	$\tilde{\Delta}_0$
1	20.4258	0.3402	40.9927	0.9687	0.2017	0.04
5	30.6721	0.1216	24.9403	0.9812	0.0165	0.04
10	31.7490	0.0900	21.9999	0.9910	0.0089	0.0176
$\hat{\mathcal{P}}_{\mathrm{PU}} \approx 15$	32.0381	0.0800	21.0000	0.9950	0.0071	0.0092
25	32.6405	0.0700	20.9996	0.9970	0.0038	0.0105
75	32.7804	0.0500	19.0121	0.9989	0.0019	0.0067

TABLE 2.3: Optimized Solutions for $\Delta_{\mathrm{T}} = 0.04$, $\mathcal{P}_{\mathrm{SU}} = 0.1$, Energy Detector

$\mathcal{P}_{\mathrm{PU}}$	max I_{SU}	$\tilde{\tau}_0$	$\tilde{\lambda}_0$	$\tilde{P}_{\mathrm{D0}}$	$\tilde{P}_{\mathrm{F0}}$	$\tilde{\Delta}_0$
0.1	2.0473	0.0938	9.1668	0.5167	0.4472	0.04
2	1.7302	0.0257	9.5706	0.2925	0.0153	0.04
$\hat{\mathcal{P}}_{\mathrm{PU}} \approx 4.1$	1.3386	0.0114	13.2715	0.1350	0.00036	0.04
8	1.2250	0.1100	23.0000	0.9950	0.0177	0.00066
16	1.2820	0.1000	26.000	0.999	0.0037	0.000035

1. When $\mathcal{P}_{\mathrm{PU}} > \hat{\mathcal{P}}_{\mathrm{PU}}$, the SU performance curve is in the High Interference Regime for SU where $\tilde{I}_{\mathrm{SU-N}}$ increases with increase in $\tilde{P}_{\mathrm{D}}$ for the non-EDI region as shown in Figure 2.4. In this case, maximum value of $\tilde{I}_{\mathrm{SU-N}}$ is at the maximum allowed value of $\tilde{P}_{\mathrm{D}}$, i.e., $P_{\mathrm{D_{T2}}}$. We cannot have a better detection probability, because, beyond $\tilde{P}_{\mathrm{D}} = P_{\mathrm{D_{T2}}}$ the EDI region begins, and in this region the SU performance decreases with increase in detection probability, as discussed in Section 2.5.1. Thus we conclude that, for a practical system where the PU transmission-power level is very high, the SU's signal detector should be designed for the maximum allowed detection probability in order to avoid any chance of interference from PU.

2. Note that for the High Interference Regime for SU, our algorithm gives us the value of detection probability at which the EDI region begins, i.e., $P_{\mathrm{D_{T2}}}$, since it is the detection probability corresponding to the optimal solution for this regime, according to remark 1.

3. When $\mathcal{P}_{\mathrm{PU}} < \hat{\mathcal{P}}_{\mathrm{PU}}$, the SU performance curve is in the Low Interference Regime for SU. For this regime, the interference caused by PU is so low that the contention occurring due to

the decrease in detection probability does not deteriorate SU performance much. Rather, it causes a decrease in sensing time which implies an increase in degrees of freedom for communications. Consequently, this results in an increase in SU information rate with decrease in detection probability. Below a certain detection probability, the SU performance for the weaker low interference regime even exceeds ν_1, i.e., the upper bound on SU performance for the high interference regime. Thus, for the *Weaker Low Interference Regime* $\tilde{I}_{\mathrm{SU-N}}$ decreases with increase in $\tilde{P}_{\mathrm{D}}$ for the non-EDI region as shown in Figure 2.4. Therefore in this regime, the maximum value of $\tilde{I}_{\mathrm{SU-N}}$ is at the minimum allowed $\tilde{P}_{\mathrm{D}}$, i.e., $P_{\mathrm{D_T}}$. We cannot have a further lower detection probability, because, by definition (2.15), $P_{\mathrm{D_T}}$ is the value of detection probability at which the PU performance degradation factor Δ_{PU} is at its maximum allowed value Δ_{T}. Further decreasing the detection probability increases Δ_{PU} which violates our optimality constraint.

4. According to remark 3, the value of PU performance degradation factor corresponding to the optimal solution for the Weaker Low Interference Regime is equal the maximum allowed value Δ_{T}. Thus, $\tilde{\Delta}_0 = \Delta_{\mathrm{T}}$ for the Weaker Low Interference Regime, as indicated by the solutions presented in Table 2.2 and 2.3. Therefore we can also infer that, for a practical system, the tolerance-level for the PU performance-degradation can be exploited to increase the SU performance by sacrificing P_{D} but only for the *Weaker Low Interference Regime of SU*.

5. Table 2.2 shows that for SU transmission power, $\mathcal{P}_{\mathrm{SU}} = 10$, the optimal SU information rate is quite high in case of high PU transmission power provided the detection probability is very high. Thus for the assumed values of the system parameters we can conclude that if 4% performance degradation can be tolerated by PU then in case of strong SU transmission power level, e.g., 10, SU has the best optimal-performance in case of High Interference Regime provided the detection probability is very high. This is because for high SU trans-

mission power the detection probability should be as much high as possible, i.e., nearly 1, in order to avoid a high interference in PU system. Also, for very high detection probability, high PU transmission power ensures smaller sensing time for SU thereby causing high degrees of freedom for communications and hence high information rate for SU.

6. Table 2.3 shows that for a weak SU transmission power, i.e., $\mathcal{P}_{\mathrm{SU}} = 0.1$, the optimal SU information rate is quite high in case of low PU transmission power. This is because, a low SU transmission power level causes a low interference to the PU system. In this case PU can tolerate a large extent of contention. This allows SU to operate at such low detection probability levels such that the sensing time is small which implies high degrees-of-freedom for communications for SU. This results in a higher SU performance for the weaker low interference regime which is even higher than the upper bound on SU performance for high interference regime.

7. According to remarks 5 and 6, for a practical system, the optimal secondary user throughput is high when both SU and PU transmission-power levels are either low or high.

2.5.4 Comparison with Conventional Studies

Recall from Section 2.4 that in the conventional studies, a major consequence of the assumption that SU-RX is coherent to PU-TX is that the inequality constraint in (4.31), on the detection probability, is always equivalent to the equality constraint, regardless of the interference caused by PU to SU. However, as discussed in Section 2.5.3, our results demonstrate that under the assumption that SU-RX is incoherent to PU-TX, the SU system can improve its throughput by sacrificing the detection probability and maintaining only the minimum required detection quality, but for the low interference regime. Thus, under the incoherent assumption, the inequality constraint on the detection probability is equivalent to the equality constraint only for this regime. Note that this does not hold for the high interference regime where the detection probability cannot be sacrificed, as discussed in Section 2.5.3.

2.6 Nearest Neighbor Decoding for PU

In the case of a Gaussian channel, the capacity, i.e., $1/2\log(1 + \text{SNR})$ can be achieved for Gaussian channel input and the *nearest neighbor decoding* (NND) is the optimal decoding rule for the receiver [33]. Since the PU system does not know about the existence of the SU system therefore we can assume that it anticipates the channel as Gaussian. For this, PU-TX uses Gaussian channel-input and NND is the decoding rule for PU-RX. For an ergodic non-Gaussian channel, the achievable channel capacity is $1/2\log(1 + \text{SNR})$ if the receiver uses NND [31]. Also, this capacity is achieved for the Gaussian channel-input [31]. Therefore, in our new scenario for PU in which PU-TX still uses the Gaussian channel-input but PU-RX uses the nearest neighbor decoding rather than the ML, the mutual information for a single channel use is $\hat{I}_{\mathrm{PU-Nc}} = 1/2\log[1 + (\mathcal{P}_{\mathrm{PU}}/\sigma^2_{\mathrm{Z_{PU}}})]$ during the communication period of SU, where $\sigma^2_{\mathrm{Z_{PU}}} = 1 + (1 - P_{\mathrm{D}})\mathcal{P}_{\mathrm{SU}}$. The mutual information for a single channel use during the sensing period of SU, $\hat{I}_{\mathrm{PU-Ns}}$, is the same as before. Therefore, the mutual information for the PU system over N channel uses can be evaluated as follows by using (2.3).

$$I_{\mathrm{PU-N}} = \pi_1 N\frac{\tau}{2}\log(1 + \mathcal{P}_{\mathrm{PU}}) + \pi_1 N\frac{(1-\tau)}{2}\log(1 + \frac{\mathcal{P}_{\mathrm{PU}}}{1 + (1 - P_{\mathrm{D}})\mathcal{P}_{\mathrm{SU}}}). \tag{2.20}$$

Thus, when PU-RX employs NND, only the expression for $I_{\mathrm{PU-N}}$ changes. The rest of the details remains the same as that in case of ML-detector for PU-RX. We can find the new optimal sensing time and threshold for SU-TX by using the updated expression for $I_{\mathrm{PU-N}}$ given by (2.20) and adopting the same methodology discussed earlier.

2.6.1 Effect of Fading

In this section, we discuss a simple scenario that can give us some idea how fading can affect the degrees of freedom and interference trade-off. However, a more thorough analysis with fading needs a dedicated study and is left for future work. For this study, we assume that PU-RX uses NND decoding and fading only occurs over the SU-TX to PU-RX channel. Let Φ_{SP} be a random variable representing the fading coefficient over this channel, such that the fading deteriorates the

SU interference to PU, i.e., $\theta_{\rm SP} = \mathrm{E}\left[|\Phi_{\rm SP}|^2\right] \in (0,1)$. Note that E[.] represents the expected value of the random variable. The PU received signal for this scenario will be $Y_{\rm PU} = X_{\rm PU} + Z_{\rm PU}$, where the noise at the PU receiver is $Z_{\rm PU} = W_{\rm PU} + \alpha\Phi_{\rm SP}X_{\rm SU}$. Similar to (2.20), the PU information rate under the stated fading scenario and NND decoding is given by, [31],

$$I_{\rm PU-N} = \pi_1 N\frac{\tau}{2}\log(1+\mathcal{P}_{\rm PU}) + \pi_1 N\frac{(1-\tau)}{2}\log(1+\frac{\mathcal{P}_{\rm PU}}{1+(1-P_{\rm D})\theta_{\rm SP}\mathcal{P}_{\rm SU}}). \tag{2.21}$$

Note that here we assume that $\Phi_{\rm SP}$ is independent of all other random variables. Since $0 < \theta_{\rm SP} < 1$, therefore, comparing (2.20) and (2.21) reveals that the PU performance improves under the fading over SU-TX to PU-RX channel that deteriorates the SU interference to PU. Therefore, for a fixed minimum allowed degradation in PU performance, PU can tolerate more interference from SU under such fading scenario. Thus, SU can further improve its performance under such scenario by sacrificing the detection probability further. However, as discussed in Section 2.5.3, such improvement in SU performance is only possible under the low interference regime.

2.7 Energy Detector versus Envelope Detector

Earlier we assumed that SU-TX uses the *optimal signal detector* under the Neyman-Pearson rule [34], i.e., the energy detector in our system model under the Gaussian assumption. In this section, we numerically solve the same optimization problem for a relatively less sensitive *envelope detector*. For the envelope detector, instead of evaluating the closed form expressions for $P_{\rm D}$ and $P_{\rm F}$, we adopt the *Monte Carlo* simulation approach. The SU and PU performances versus the detection probability for both the detectors are presented in Figures 2.6 and 2.7 respectively. A particular optimal solution for the envelope detector is presented in Table 2.4.

As compared to Table 2.2, Table 2.4 shows that the optimal sensing time is almost the same as in case of the optimal (energy) detector and the most significant effect of using the envelope detector instead of the energy detector is a considerable reduction in the value of optimal sensing threshold $\tilde{\lambda}_0$. This is because the envelope detector is less sensitive to the PU signal than the energy detector. To compensate this decrease in sensitivity, the sensing threshold needs to be

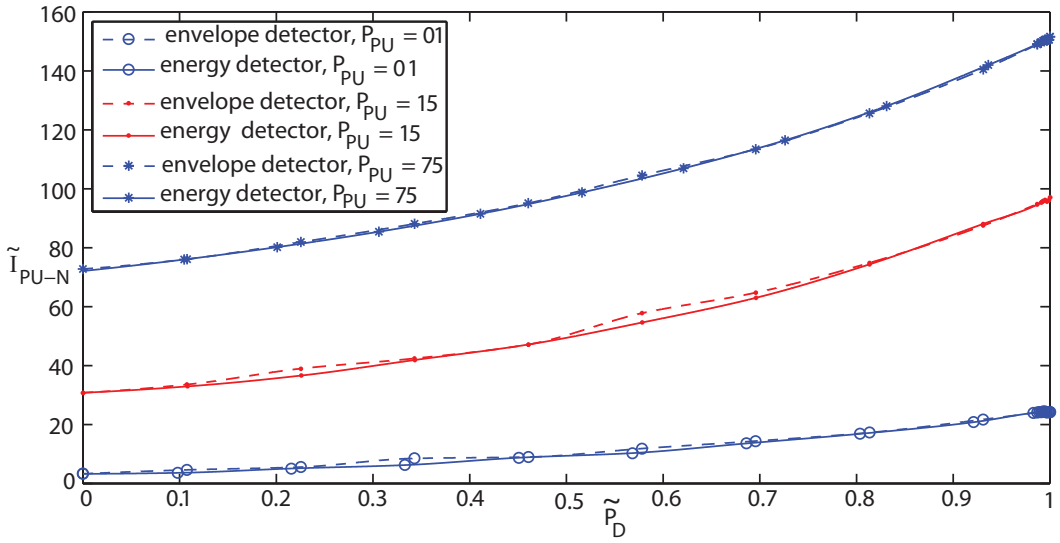


FIGURE 2.6: PU performance curves for envelope and energy detectors.

lowered while keeping the sensing time the same as in case of the optimal (energy) detector. Also, this reduction in the sensing threshold increases the false alarm probability for the envelope detector as compared to the energy detector while the sensing time remains the same.

TABLE 2.4: Optimized Solutions for $\Delta_{\mathrm{T}} = 0.04$, $\mathcal{P}_{\mathrm{SU}} = 10$, Envelope Detector

$\mathcal{P}_{\mathrm{PU}}$	max I_{SU}	$\tilde{\tau}_0$	$\tilde{\lambda}_0$	$\tilde{P}_{\mathrm{D0}}$	$\tilde{P}_{\mathrm{F0}}$	$\tilde{\Delta}_0$
1	17.8873	0.3059	21.3106	0.9687	0.3441	0.0398
$\hat{\mathcal{P}}_{\mathrm{PU}} \approx 15$	31.8157	0.0800	11.0000	0.9930	0.0073	0.0135
75	32.6701	0.0500	8.0000	0.9990	0.0047	0.0068

Figure 2.6 shows that there is no significant difference in PU performances for both the detectors. On the other hand, Figure 2.7 shows that in case of the high interference regime there is almost no difference in SU performances for both the detectors but in case of the low interference regime SU performance for the envelope detector is slightly less than that for the energy detector. This is because in the high interference regime PU transmission power is very high, and together with the appropriate decrease in the sensing threshold, it almost completely compensates the decrease in detector's sensitivity.

Frequency bands dedicated for TV-broadcasting are the potential spectral resource that can be used by the intelligent wireless devices (SUs) that can sense the presence or absence of the TV

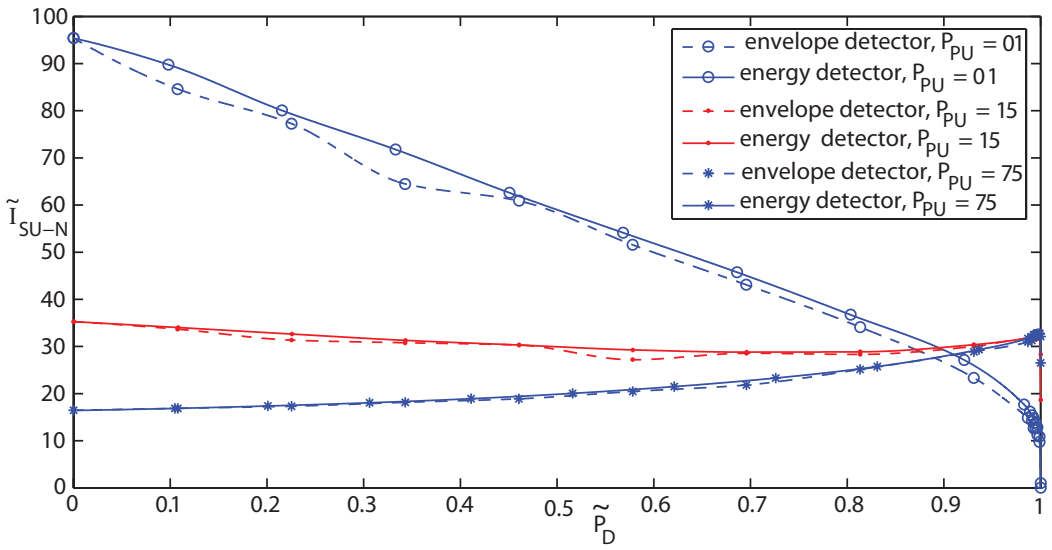


FIGURE 2.7: SU performance curves for envelope and energy detectors.

signal in the frequency band of interest and access the channel if it is found idle [35]. Since TV broadcasting is done at a high transmission power level therefore, according to our results in this section, even if the signal detector of the SU wireless device performs suboptimally SU can still achieve the optimal-detector's performance level just by adjusting the sensing parameters, e.g., in the same way as the sensing threshold for the envelope detector is tuned to achieve the optimal detector's performance level in the high interference regime, in our system model.

2.8 Conclusion

In this study, we explored the degrees-of-freedom and interference trade-off for a simple cognitive radio system under an assumption that the secondary-user (SU) receiver operates incoherently with the primary-user (PU) transmitter. This assumption makes our system model practically more rigorous, as compared to the conventional studies where a coherent operation is assumed. Due to the incoherent operation of SU receiver in our model, the resulting interference channel is mixture-Gaussian. Our objective is to find the optimal sensing threshold and sensing time for the signal detector that is used by the SU to detect the channel occupancy by the PU. The optimality criterion is to maximize the SU performance and at the same time keep the PU performance degradation, caused by the interference due to SU, within a tolerable range. We numerically solve

this optimization design-problem since the closed form solution is not possible. As a further novel contribution, we define the interference regimes for SU performances on the basis of PU transmission power level. We also find that the tolerance level for PU performance degradation can be exploited to increase the SU performance but only for the *Weaker Low Interference Regime for SU*. The effect of a more robust decoding strategy, namely, nearest neighbor decoding approach, at PU-RX is also elaborated. Finally, we demonstrate that even if the signal detector of SU performs suboptimally, SU can still achieve the optimal detector's performance level in the high interference regime just by adjusting the sensing parameters accordingly.

Chapter 3

Communication Channel Allocation by Call Admission and Preemption Control

Public safety communication systems are implemented as trunked mobile radio systems. The finite-source queueing model, i.e., the Engset model, has been suggested as the appropriate model in literature for these systems. In this study, we present the analytical framework and characteristics of the optimal call admission and preemption control for such systems. We consider two classes of users, namely, of high and low priority. We show that the optimal low priority (LP) call admission policy is a state dependent threshold based policy. We also demonstrate that the optimal LP call preemption policy is a threshold based policy. This study has resulted in two research papers, [36, 37].

3.1 Introduction

Public safety communication systems are implemented as trunked mobile radio systems [38–40]. In these systems, the coverage area is divided into *sites*. Also, the users are divided into fleets or groups normally called *talk groups*. Each user can only communicate with another user of the same talk group. When a user needs to talk, it presses the push-to-talk (PTT) button of its radio, in order to send a call request at the site's *base-station*. The base-station assigns communication resources to the caller and the rest of the users of the talk group, in order to broadcast the voice call throughout the group. Thus, in these systems, the channel allocation for communication is performed on talk group basis. Therefore, instead of radios the source units are the talk groups. These talk groups are also finite in number. In these systems, when a user in a talk group talks, the rest of the users only listen. Therefore, a talk-group under service cannot generate any call. Thus, only an idle talk-group can generate a call in a trunked radio system. This makes the system call arrival rate depend on the number of idle talk groups in the system, which in turn depends on the system state, defined as the number of busy talk groups in the system. These features of the

trunked systems, namely, the finite number of talk groups and the state dependent call arrival rates, are well captured by the finite-source or the *Engset model* [41], discussed in detail in Section 3.2. Therefore the finite-source model is suggested for the trunked radio systems in [39, 40]. This traffic model is different from the Poisson arrival traffic in the *Erlang system*, where the system call-arrival rate is state independent [41].

Due to the limited number of channels available, trunking alone is not sufficient to fulfill the service demand of the system, especially in high call-traffic scenarios, e.g., an emergency or a disaster event in case of the first responders' communication systems. To alleviate this problem talk groups are assigned priority levels, such that the system serves the talk groups according to the assigned priority, so that whenever there is a high demand of service the high priority talk groups get least affected. We call this system as a *prioritized trunked radio system*, i.e., a system which consists of talk groups of different priority levels. An efficient bandwidth allocation and management mechanism is required for any prioritized communication system, which has a limited number of communication resources. This goal can be accomplished by introducing call admission and preemption mechanisms [42–44]. To the best of our knowledge, the study of optimal admission and (or) preemption control for the queueing systems has only been done for the Poisson arrival traffic, i.e., for the *Erlang system*, see [42, 45–47] and references therein. In this chapter, we study the admission and preemption control for an Engset (finite source) system, which is the suggested model for the trunked systems [39, 40].

Our analysis of finite source system is non-trivial in spite of the extensive literature on relevant analysis for the system with Poisson arrivals, i.e., the Erlang system. This is because the analysis for the Erlang system cannot be immediately extended to the finite source system due to two fundamental reasons: 1. firstly, the net call arrival rate in the finite source system depends on the system state [41], in contrast to the Poisson arrival traffic in the Erlang system where the call arrival rate is state independent [41]; and 2. secondly, the topology of the state space of the finite source system is different from the Erlang system, as demonstrated in Section 3.2.1. These

aspects of the finite source system also make the analysis more challenging as compared to that of the Erlang system. More specifically, our major contributions are: 1. we present the analytical framework for the optimal control of both admission and preemption in a two-class finite source loss system; and 2. we show that the optimal low priority (LP) call admission policy is a state dependent threshold based policy, which is similar to the result for the optimal admission control in Erlang system presented in [46]. Also as a further contribution, we provide the sufficient cost conditions for which the LP call admission and preemption are *inevitable*, i.e., always optimal, independent of the system state. Finally, we demonstrate using numerical results that the optimal LP call preemption policy in a finite source loss system is a threshold based policy.

More Applications: In a broader context, in this study, we investigate the optimal service allocation for a finite-source queueing system. Apart from trunked radios, this model, and hence our results, are also applicable to other interesting systems. For example, consider an *Internet of things*, similar to one considered in [48]. In such systems, sensor and actuator devices are connected to Internet via a wireless access point. We can model this as a finite source queueing system, with devices as source units and the access point as a channel, such that a source-unit under service does not generate a new service request. This is exactly the model being considered in this study. Thus, our study can help develop optimal service allocation policy for such systems as well.

3.2 Mathematical Model

We consider a finite source model for a trunked radio system where the talk groups are the call source units, as discussed in Section 4.1. We consider a *loss* system, i.e., a system without a queueing buffer or zero queue length. We assume two classes of talk groups, namely, the low-priority (LP) and the high-priority (HP) talk groups. Let M_{L} and M_{H} be the number of LP and HP talk groups in the system respectively. There are N identical channels in the system. We also assume that the total number of talk groups in the system is more than the number of channels, i.e., $M_{\mathrm{L}} + M_{\mathrm{H}} > N$. The call service times, or call durations, are independent and exponentially

distributed, with mean μ^{-1}, for both the classes. Thus μ is the service completion rate for both LP and HP calls. We consider the same service rates for both classes as done in [42, 47]. Also, only an idle talk group can generate a call request as explained in Section 4.1. This makes the net call arrival rate in the system depend on the *system state*. We define the *system state* as the number of busy channels in the system, that is also equal to the number of busy talk groups. In this study, the idle times of the talk groups are independent and exponentially distributed. Let λ_{L} and λ_{H} be the LP and HP call arrival rates, per idle source, respectively. For a finite-source loss system, same as the one we are considering in this study, the system steady-state (long-run) average performance is insensitive to the idle and service times' distributions of the source unit [41]. We consider exponential idle and service times for the talk groups, so that the already established Markov Decision Process (MDP) theory for the exponential queueing systems can be used, which is presented in detail in Section 3.3.

Decision Control: In this study we are interested in the characteristics of the optimal call admission and preemption control for the system under consideration. At any given time, the system sees one and only one arrival, due to the assumption of independent exponentially distributed talk group idle times [41]. This call can either be an HP or an LP call. We then have two cases. The first case arises when the arrived call is HP. In this case, the arrived HP call is always admitted if an idle channel is available. Otherwise, if all channels are busy, then the system needs to decide out of two options, i.e., either block the arrived HP call or permanently remove an LP call from the system if there is one already under service. The latter option is known as *preemption* and this decision process is called the *Low-Priority Preemption Control* (LPPC). The second case arises when the arrived call is LP. In this case, the arrived LP call is always blocked if all channels are busy. Otherwise, if an idle channel is available, the system needs to decide out of two options, i.e., either block the arrived LP call or admit it. This decision process is called the *Low-Priority Admission Control* (LPAC). Thus, there are two decision controls that we are considering jointly,

namely, LPPC and LPAC. Together we name the decision control as the *LP Call Admission and Preemption (LCAP) Control*.

3.2.1 System State Space

We define the state of the system in the form of a tuple $(n_{\mathrm{L}}, n_{\mathrm{H}})$, $0 \leq n_{\mathrm{L}} \leq \hat{M}_{\mathrm{L}}$, $0 \leq n_{\mathrm{H}} \leq \hat{M}_{\mathrm{H}}$. Here n_{L} and n_{H} are the numbers of LP and HP calls that are under service in the system, respectively, and $\hat{M}_{\mathrm{L}}$ and $\hat{M}_{\mathrm{H}}$ denote the maximum number of LP and HP calls that can be in service in the system at a given time. Mathematically, $\hat{M}_{\mathrm{L}} = \min(M_{\mathrm{L}}, N)$ and $\hat{M}_{\mathrm{H}} = \min(M_{\mathrm{H}}, N)$. In case of Poisson traffic [47], the number of LP or HP calls under service can only be less than or, at the most, equal to the number of channels N, i.e., $\hat{M}_{\mathrm{L}} = \hat{M}_{\mathrm{H}} = N$. So the system state space, in case of Poisson traffic, is always in the form of a 2D triangular grid [47]. However, in our finite source system model, the system state space can be either a 2D triangular grid, when $M_{\mathrm{L}} \geq N$ and $M_{\mathrm{H}} \geq N$, or pentagonal, as shown in Figures 3.1(a) and 3.1(c), when $M_{\mathrm{L}} < N$ and $M_{\mathrm{H}} < N$, or tetragonal grid otherwise.

We define $\mathcal{S}$ as the system state space, i.e.:

$$\mathcal{S} = \{(n_{\mathrm{L}}, n_{\mathrm{H}}) \;\big|\; 0 \leq n_{\mathrm{L}} \leq \hat{M}_{\mathrm{L}}, 0 \leq n_{\mathrm{H}} \leq \hat{M}_{\mathrm{H}},\ n_{\mathrm{L}} + n_{\mathrm{H}} \leq N\}. \tag{3.1}$$

We now present a classification of states on the basis of the LCAP decision control described in Section 3.2. This will help analyze each group of states that is characterized by a certain decision control as explained below.

LPAC States (G1): Let $\mathcal{S}_1 \subset \mathcal{S}$, defined as:

$$\mathcal{S}_1 = \{(n_{\mathrm{L}}, n_{\mathrm{H}}) \;\big|\; 0 \leq n_{\mathrm{L}} \leq \hat{M}_{\mathrm{L}} - 1, 0 \leq n_{\mathrm{H}} \leq \hat{M}_{\mathrm{H}},\ n_{\mathrm{L}} + n_{\mathrm{H}} \leq N - 1\}. \tag{3.2}$$

These states are marked with crosses and circles in Figure 3.1(a) and only with crosses in Figure 3.1(c). Whenever the system is in any of the states in $\mathcal{S}_1$, there is always an idle channel available in the system. Therefore, all the states in $\mathcal{S}_1$ require only the LPAC control that is presented in Section 3.2.

LPPC States (G2): Let $\mathcal{S}_2 \subset \mathcal{S}$, defined as:

$$\mathcal{S}_2 = \{(n_{\mathrm{L}}, n_{\mathrm{H}}) \mid N - \hat{M}_{\mathrm{H}} + 1 \leq n_{\mathrm{L}} \leq \hat{M}_{\mathrm{L}}, n_{\mathrm{L}} + n_{\mathrm{H}} = N\}. \tag{3.3}$$

These states are marked with asterisks and squares in Figures 3.1(a), and only with asterisks in Figure 3.1(c). Whenever the system is in any of the states in $\mathcal{S}_2$, all the channels are busy in the system. Therefore, all the states in $\mathcal{S}_2$ require only the LPPC control that is presented in Section 3.2.

Known-LPAC State (G3): Let $\mathcal{S}_3 \subset \mathcal{S}$, defined as:

$$\mathcal{S}_3 = \{(n_{\mathrm{L}}, n_{\mathrm{H}}) \mid n_{\mathrm{L}} = N - \hat{M}_{\mathrm{H}},\ n_{\mathrm{H}} = \hat{M}_{\mathrm{H}}\}. \tag{3.4}$$

This group only consists of a single state. Example of this state is (10, 30) in Figures 3.1(a) and 3.1(c). This state does not require any LPPC control, but has a known or fixed LPAC policy, i.e., whenever an LP call arrives it will always be blocked, according to our decision policy defined in Section 3.2.

No-Policy States (G4): Let $\mathcal{S}_4 \subset \mathcal{S}$, defined as:

$$\mathcal{S}_4 = \{(n_{\mathrm{L}}, n_{\mathrm{H}}) \mid n_{\mathrm{L}} = \hat{M}_{\mathrm{L}},\ 0 \leq n_{\mathrm{H}} \leq N - \hat{M}_{\mathrm{L}} - 1\}. \tag{3.5}$$

This group of states exist when $M_{\mathrm{L}} < N$. These states are marked with diamonds for $n_{\mathrm{L}} = 25$ in Figures 3.1(a) and 3.1(c). These states neither require an LPAC, as all LP talk groups are busy in these states, nor they require an LPPC, as an idle channel is always available for a new HP call.

3.2.2 The MDP Approach

Given the stationary LCAP control policy, the state of the system evolves as a 2D continuous-time *Markov Decision Process* (MDP) [49]. The MDP parameters of our model are discussed below.

Rewards and Costs: In this study, we do not assign any rewards for call admission or service completion, rather we only assign costs for blocked and preempted calls. Let constants $A_{\mathrm{L}} > 0$ and $A_{\mathrm{H}} > 0$ be the costs of blocking LP and HP calls respectively. It is reasonable to assume

$A_{\mathrm{L}} \leq A_{\mathrm{H}}$, due to the priority assignment of the calls. Let $K > 0$ be the cost of preempting an LP call which is already under service in the system. A similar cost model is also used in [42].

Discounting: We use discounting at a rate $\alpha \geq 0$. This ensures that the costs, at time, say t, are scaled by a factor of $\exp(-\alpha t)$. Note that the equivalence of α discounting and an exponential lifetime of the system with rate α is well-known [50]. See [46, 47] for more details.

Uniformization: The process we have is a continuous-time Markov chain. We develop the discrete time equivalent of this system using uniformization technique [51]. Without loss of generality, we set the maximum possible rate out of a state to 1, i.e., $M_{\mathrm{L}}\lambda_{\mathrm{L}}+M_{\mathrm{H}}\lambda_{\mathrm{H}}+N\mu+\alpha = 1$. Hence, an LP call arrives with probability $(M_{\mathrm{L}}-n_{\mathrm{L}})\lambda_{\mathrm{L}}$, an HP call arrives with probability $(M_{\mathrm{H}}-n_{\mathrm{H}})\lambda_{\mathrm{H}}$, an LP call completes its service with probability μn_{L}, an HP call completes its service with probability μn_{H}, the process terminates with probability α and the system stays at the same state, due to no arrival and no service completion, with probability $n_{\mathrm{L}}\lambda_{\mathrm{L}} + n_{\mathrm{H}}\lambda_{\mathrm{H}} + (N - n_{\mathrm{L}} - n_{\mathrm{H}})\mu$.

Criterion for Policy Optimization: The objective is to find an optimal LCAP control policy that minimizes the total expected long-run average cost for the system. However, we formulate and study the problem for the expected α-discounted cost over the finite horizon, in order to utilize useful analysis techniques such as induction, as used in [47]. Similarly as claimed in [47], our results can also be extended to the long run or infinite horizon case, according to the conditions discussed in [52]. Now, consider the average cost case that corresponds to $\alpha \to 0$ [46, 47]. Since, our decision control or action space and the system state space are finite, therefore the results for the the α-discounted case also hold for the average cost case [46, 47].

3.3 Dynamic Programming Formulation

In this section, we formulate our finite-horizon discounted-cost MDP-problem, as a *stochastic dynamic programming* (DP) problem [53]. Let m be the observation points left until the end of the horizon. Note that it is a reversed time index, i.e., $m = 0$ denotes the end of the horizon. We define the cost function, $C_m(n_{\mathrm{L}}, n_{\mathrm{H}})$, as the minimum expected discounted cost for the system in the current state $(n_{\mathrm{L}}, n_{\mathrm{H}})$ at time period m. We assume that $C_0(n_{\mathrm{L}}, n_{\mathrm{H}}) = 0,\ \forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}$,

same as in [42, 47]. This assumption means that the system closes without paying any cost at the end of the process horizon. Now, we present the DP equations for $m > 0$, for each group of states described in Section 3.2.1. Here we adopt a similar approach to write these equations as done in [42, 46, 47].

For LPAC States: For these states, the optimal LPAC policy decision in state $(n_{\mathrm{L}}, n_{\mathrm{H}})$ at time m is as follows:

$$\hat{d}_{\mathrm{LPAC}} = \begin{cases} +1, \ C_{m-1}(n_{\mathrm{L}}+1, n_{\mathrm{H}}) \leq A_{\mathrm{L}} + C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}}) \\ -1, \ \text{otherwise.} \end{cases} \tag{3.6}$$

Here, $+1$ and -1 mean "admit LP call" and "block LP call", respectively. Now for these states, i.e., $\forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_1$, the DP equation for $m > 0$ is:

$$\begin{aligned} C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) = {} & (M_{\mathrm{L}} - n_{\mathrm{L}})\lambda_{\mathrm{L}} \times \min\{A_{\mathrm{L}} + C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}}), C_{m-1}(n_{\mathrm{L}}+1, n_{\mathrm{H}})\} \\ & + (M_{\mathrm{H}} - n_{\mathrm{H}})\lambda_{\mathrm{H}} C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}}+1) \\ & + n_{\mathrm{L}}\mu C_{m-1}(n_{\mathrm{L}}-1, n_{\mathrm{H}}) + n_{\mathrm{H}}\mu C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}}-1) \\ & + \left[n_{\mathrm{L}}\lambda_{\mathrm{L}} + n_{\mathrm{H}}\lambda_{\mathrm{H}} + (N - n_{\mathrm{L}} - n_{\mathrm{H}})\,\mu\right] C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}}). \end{aligned} \tag{3.7}$$

The first term is the contribution to the cost due to an LP arrival. Here, we have the option to either block the LP call by paying a cost A_{L} or admit it without any cost. The second term is the contribution to the cost due to an HP arrival. We always admit an arriving HP call in these states according to our policy structure mentioned in Section 3.2, because there is always an idle channel available in these states as discussed in Section 3.2.1. The third and fourth terms are the contributions to the cost due to service completions. The fifth term is a contribution to the cost if the system remains in the same state due to no arrival and no service completion. The probabilities corresponding to all these events are mentioned in Section 3.2.2. Similarly, we write the equations for the rest of the groups of states.

For LPPC States: For these states, the optimal LPPC policy decision in state $(n_{\mathrm{L}}, n_{\mathrm{H}})$ at time m is as follows:

$$\hat{d}_{\mathrm{LPPC}} = \begin{cases} +2, \ K + C_{m-1}(n_{\mathrm{L}} - 1, n_{\mathrm{H}} + 1) \leq A_{\mathrm{H}} + C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}}) \\ -2, \ \text{otherwise.} \end{cases} \tag{3.8}$$

Here, $+2$ and -2 mean "admit HP call by preempting an LP call" and "block HP call", respectively. Now for these states, i.e., $\forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_2$, the DP equation for $m > 0$ is:

$$\begin{aligned} C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) &= (M_{\mathrm{L}} - n_{\mathrm{L}})\lambda_{\mathrm{L}} \left[A_{\mathrm{L}} + C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}})\right] \\ &\quad + (M_{\mathrm{H}} - n_{\mathrm{H}})\lambda_{\mathrm{H}} \min\{A_{\mathrm{H}} + C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}}), K + C_{m-1}(n_{\mathrm{L}} - 1, n_{\mathrm{H}} + 1)\} \\ &\quad + n_{\mathrm{L}}\mu C_{m-1}(n_{\mathrm{L}} - 1, n_{\mathrm{H}}) + n_{\mathrm{H}}\mu C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}} - 1) \\ &\quad + \left[n_{\mathrm{L}}\lambda_{\mathrm{L}} + n_{\mathrm{H}}\lambda_{\mathrm{H}}\right] C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}}). \end{aligned} \tag{3.9}$$

The first term is the contribution to the cost due to an LP arrival. We always block an LP call by paying a cost A_{L} in these states, according to our policy structure mentioned in Section 4.1, because all channels are busy in these states as discussed in Section 3.2.1. The second term is the contribution to the cost due to an HP arrival. Here, we have the option to either block the HP call in which case a cost A_{H} is incurred or admit the HP call by preempting an LP call already in service by paying a cost K. The third and fourth terms are the contributions to the cost due to service completions. The fifth term is a contribution to the cost if the system remains in the same state due to no arrival and no service completion.

For Known-LPAC State: For $(n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_3$:

$$\begin{aligned} C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) &= (M_{\mathrm{L}} - n_{\mathrm{L}})\lambda_{\mathrm{L}} \left[A_{\mathrm{L}} + C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}})\right] \\ &\quad + (M_{\mathrm{H}} - n_{\mathrm{H}})\lambda_{\mathrm{H}} \left[A_{\mathrm{H}} + C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}})\right] \\ &\quad + n_{\mathrm{L}}\mu C_{m-1}(n_{\mathrm{L}} - 1, n_{\mathrm{H}}) + n_{\mathrm{H}}\mu C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}} - 1) \\ &\quad + \left[n_{\mathrm{L}}\lambda_{\mathrm{L}} + n_{\mathrm{H}}\lambda_{\mathrm{H}}\right] C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}}). \end{aligned} \tag{3.10}$$

The first term is the contribution to the cost due to an LP arrival. The second term is the contribution to the cost due to an HP arrival. We always block an arriving call in this state, by paying a cost A_{L} for an LP arrival and A_{H} for an HP arrival, as discussed in Section 3.2.1. The third and fourth terms are the contributions to the cost due to service completions. The fifth term is a contribution to the cost if the system remains in the same state due to no arrival and no service completion.

For No-Policy States: $\forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_4$:

$$\begin{aligned} C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) = {} & (M_{\mathrm{H}} - n_{\mathrm{H}})\lambda_{\mathrm{H}} C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}} + 1) \\ & + n_{\mathrm{L}}\mu C_{m-1}(n_{\mathrm{L}} - 1, n_{\mathrm{H}}) + n_{\mathrm{H}}\mu C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}} - 1) \\ & + \left[n_{\mathrm{L}}\lambda_{\mathrm{L}} + n_{\mathrm{H}}\lambda_{\mathrm{H}} + (N - n_{\mathrm{L}} - n_{\mathrm{H}})\,\mu\right] C_{m-1}(n_{\mathrm{L}}, n_{\mathrm{H}}). \end{aligned} \tag{3.11}$$

The first term is the contribution to the cost due to an HP arrival. As discussed in Section 3.2.1, for these states we cannot have an LP arrival, rather we can only have an HP call arrival which is always admitted. The second and third terms are the contributions to the cost due to service completions. The fourth term is a contribution to the cost if the system remains in the same state due to no arrival and no service completion.

Boundary Conditions: The following boundary conditions are set, $\forall m \geq 0,\ (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}$:

$$C_m(-1, n_{\mathrm{H}}) = C_m(0, n_{\mathrm{H}}) \qquad C_m(n_{\mathrm{L}}, -1) = C_m(n_{\mathrm{L}}, 0)$$

$$C_m(n_{\mathrm{L}}, \hat{M}_{\mathrm{H}} + 1) = C_m(n_{\mathrm{L}}, \hat{M}_{\mathrm{H}}) \qquad C_m(\hat{M}_{\mathrm{L}} + 1, n_{\mathrm{H}}) = C_m(\hat{M}_{\mathrm{L}}, n_{\mathrm{H}})$$

$$C_m(n_{\mathrm{L}}, N - n_{\mathrm{L}} + 1) = C_m(n_{\mathrm{L}}, N - n_{\mathrm{L}})$$

This concludes the DP formulation of our MDP problem.

3.4 Characterization of Admission and Preemption Control

In this section, we present interesting properties of the optimal LPAC and LPPC policies. In the proofs we employ the tools that are usually used for analyzing the MDP based queueing systems, namely, the *sample path argument*, as in [46, 47], and the method of *induction*, as in [47].

3.4.1 Coupling

In the proofs, based on the sample path argument, we use the method of *coupling*, and compare two coupled systems with each other, similar to [46] where the system model assumes Poisson call arrivals. However, in our study, we have a finite-source model in which calls originate from a finite number of source units.

In case of Poisson arrivals, the busy calls of both the systems are coupled together. The coupled systems experience the same sequence of events with same properties. In particular, both the systems have the same arrival pattern, in case of Poisson arrivals. The calls in both the systems always complete their services together at the same time and leave the system if they are of the same class. However, if their classes are different, then both the calls always leave the system together if and only if both the classes have the same call service rates [46].

We now extend this concept of coupling for our finite source system. In our system, all the details of coupling remain the same as those mentioned for the Poisson arrivals, except the call arrival patterns. This is because, in a finite source system, we need to consider the coupling of the source units, including both the idle and busy sources. The calls from the coupled idle sources of the same type always arrive together at the same time for both the systems, but the calls from the coupled idle sources of different types always arrive together at the same time for both the systems if and only if both the classes have the same call arrival rates.

3.4.2 Upper Bound on Cost to Admit LP call

Here we present a lemma which will set an upper bound for the cost of an additional LP call. The first part of the lemma is used separately as Theorem 1, while the second part is used during the proof of Lemma 3.

Lemma 1. $\forall m \geq 0$, $\forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}$, *s.t.,* $(n_{\mathrm{L}}+1, n_{\mathrm{H}}) \in \mathcal{S}$*:*

1. $K \leq A_{\mathrm{L}} \implies C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \leq A_{\mathrm{L}}$.

2. $K > A_{\mathrm{L}} \implies C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \leq K$.

Proof. The methodology used for proving this lemma is the *sample path argument*. Consider two systems, A and B. Assume that system A starts in state $(n_{\mathrm{L}}, n_{\mathrm{H}})$ and B starts in state $(n_{\mathrm{L}}+1, n_{\mathrm{H}})$. We *couple* the two systems such that all the service and idle times, for the channels and sources, in both the systems are the same, except, the additional idle LP talk group and channel in system A, and the additional busy LP talk group and channel in system B. Also, assume that system A follows the optimal policy, while system B imitates all the decisions of system A. Therefore, $C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) = C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}})$ and $C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) \leq C_m^{\mathrm{B}}(n_{\mathrm{L}}+1, n_{\mathrm{H}})$, where $C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}})$ and $C_m^{\mathrm{B}}(n_{\mathrm{L}}+1, n_{\mathrm{H}})$ are the expected discounted costs of systems A and B repectively, at time m. Next, we analyze all the possible cases in which both the systems can completely couple to become identical systems.

The first case is, when both the systems end up with same costs after being completely coupled. For this case, there can only be two possible scenarios. In the first scenario, at some point, the additional idle source of system A may generate an LP call and system A admits it to couple with system B. In the second scenario, the additional busy channel of system B completes the LP call service to couple with system A. For both these scenarios:

$$C_m^{\mathrm{B}}(n_{\mathrm{L}}+1, n_{\mathrm{H}}) = C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}}). \tag{3.12}$$

The second case is, when both the systems end up with different costs after being completely coupled. For this case, there can only be two possible scenarios. In the first scenario, system A has an idle channel available and an LP call arrives. System A admits the call but system B blocks it by paying the cost A_{L}. Thus for this scenario:

$$C_m^{\mathrm{B}}(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}}) = A_{\mathrm{L}}. \tag{3.13}$$

In the second scenario, all the channels in system B are busy and an HP call arrives. Now, the only possibility that the two systems can couple completely, is when system B admits it while preempting an LP call at a cost K, provided the LP call preemption policy adopted by system B

permits this, and system A admits the call without preemption, because system A still has an idle channel available. In this scenario:

$$C_m^{\mathrm{B}}(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}}) = K. \tag{3.14}$$

From (3.12), (3.13) and (3.14), we can conclude:

$$C_m^{\mathrm{B}}(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}}) \leq \max(A_{\mathrm{L}}, K). \tag{3.15}$$

According to our assumption, $C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) = C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}})$ and $C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) \leq C_m^{\mathrm{B}}(n_{\mathrm{L}}+1, n_{\mathrm{H}})$. Therefore, we have:

$$C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \leq \max(A_{\mathrm{L}}, K). \tag{3.16}$$

This completes the proof of Lemma 1. □

3.4.3 A Sufficient Condition for Inevitability of LP Admission

The following theorem states a sufficient condition for the inevitability of LP admission, for the LPAC states.

Theorem 1. *It is always optimal to admit an arriving LP call, whenever an idle channel is available in the system,* **if** *the preemption of an LP call, at the most, costs as much as its blocking, i.e.,* $\forall m \geq 0$, $\forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_1 \subset \mathcal{S}$, *s.t.,* $(n_{\mathrm{L}}+1, n_{\mathrm{H}}) \in \mathcal{S}$*:*

$$K \leq A_{\mathrm{L}} \implies C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) \leq A_{\mathrm{L}} + C_m(n_{\mathrm{L}}, n_{\mathrm{H}}). \tag{3.17}$$

Proof. The theorem is mathematically the same as Lemma 1.1, i.e., part 1 of Lemma 1. □

We state Lemma 1.1, separately as a theorem, due to its significance, as it tells us that we only need to look for an optimal LPAC policy whenever $K > A_{\mathrm{L}}$, otherwise when $K \leq A_{\mathrm{L}}$, as shown above, the optimal policy is known for all the states, i.e., admit all arriving LP calls whenever an idle channel is available.

3.4.4 Upper Bound on Cost to Admit HP call for LPAC States

Here we present a lemma which will set an upper bound for the cost of an additional HP call, for the LPAC states. It is also used in the proof of Lemma 3.

Lemma 2. $\forall m \geq 0$, $\forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_1$, *s.t.,* $(n_{\mathrm{L}}, n_{\mathrm{H}}+1) \in \mathcal{S}$*:*

$$C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \leq A_{\mathrm{H}}. \tag{3.18}$$

Proof. The proof of this lemma is similar to that of Lemma 1, and can be found in Section 3.5.1. □

3.4.5 Convexity of Cost to Admit LP call in n_{H}

In this section we show that for a given n_{L}, the cost to add an additional LP call in the system is monotonically non-decreasing and thus convex in n_{H}, for the LPAC states. This property helps us prove Theorem 2. As mentioned in Section 3.4.3, we only need to look for an optimal LPAC policy whenever $K > A_{\mathrm{L}}$, therefore we prove Lemma 3 under this condition. Also, we only need to consider those LPAC states, i.e., $(n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_1$, such that, the states involved in the statement (3.20) of Lemma 3 belong to $\mathcal{S}$. Let the set of such LPAC states be $\mathcal{S}_5$, then:

$$\begin{aligned}\mathcal{S}_5 &= \{(n_{\mathrm{L}}, n_{\mathrm{H}}) \mid (n_{\mathrm{L}}+1, n_{\mathrm{H}}), (n_{\mathrm{L}}+1, n_{\mathrm{H}}+1), (n_{\mathrm{L}}, n_{\mathrm{H}}+1) \in \mathcal{S}\} \\ &= \{0 \leq n_{\mathrm{L}} \leq \hat{M}_{\mathrm{L}} - 1, 0 \leq n_{\mathrm{H}} \leq \hat{M}_{\mathrm{H}} - 1, n_{\mathrm{L}} + n_{\mathrm{H}} \leq N - 2\}. \end{aligned} \tag{3.19}$$

These are all the states in $\mathcal{S}_1$ excluding those for which either $n_{\mathrm{H}} = \hat{M}_{\mathrm{H}}$ or $n_{\mathrm{L}} + n_{\mathrm{H}} = N - 1$. Now we can present our next lemma, as follows:

Lemma 3. *Let* $K > A_{\mathrm{L}}$. $\forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_5 \subset \mathcal{S}_1$, $\forall m \geq 0$*:*

$$C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \leq C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1). \tag{3.20}$$

Proof. We prove this lemma by induction on m. The proof can be found in Section 3.5.2. □

3.4.6 Optimal LPAC Policy is of Threshold Type

Theorem 2. *The optimal LPAC policy is a state dependent threshold policy, i.e., for a given number of LP calls in the system n_{L}, there exists an optimal threshold $n_{\mathrm{th}}(n_{\mathrm{L}}) \in \{0, 1, \ldots, \hat{M}_{\mathrm{H}} + 1\}$, such that, if an arriving LP call finds the system in state $(n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_1$ and $n_{\mathrm{H}} < n_{\mathrm{th}}(n_{\mathrm{L}})$, the LP call is admitted. Otherwise, the LP call is blocked.*

Proof. We need to consider two cases separately, namely, $K > A_{\mathrm{L}}$ and $K \leq A_{\mathrm{L}}$.

Case 1: Let $K > A_{\mathrm{L}}$. In this case, we can use Lemma 3 which implies that an additional LP call costs the same or more in state $(n_{\mathrm{L}}, n_{\mathrm{H}} + 1) \in \mathcal{S}_1$ than it does in state $(n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_1$. Thus, for a fixed number of LP calls in the system, if it is not optimal to admit an LP call in state $(n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_1$ then it is never optimal to admit an arriving LP call for all the other states with higher number of HP calls in the system. This guarantees that the optimal policy is of threshold type, as stated in Theorem 2. A similar argument is used in [46] to show that the optimal policy is a state dependent threshold policy.

Case 2: Let $K \leq A_{\mathrm{L}}$. Here, we can use Theorem 1 which implies that, for this case LP admission is optimal for all the LPAC states. Such a policy is a special case of a threshold based LPAC policy with $n_{\mathrm{th}}(n_{\mathrm{L}}) = \hat{M}_{\mathrm{H}} + 1$, $\forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_1$. An example of this case is presented in Figure 3.1(c) and discussed in Section 3.4.10. □

This theorem helps simplify the implementation of the policy decision control for the LPAC states, because, we only need to know the n_{H}-thresholds for each n_{L}, instead of storing the decisions for all the states of the system. These thresholds can be easily determined using the well known optimal policy search algorithms, namely *value iteration* and *policy iteration* algorithms [53]. In Section 3.4.10, we present results for the policy iteration algorithm.

3.4.7 Convexity of Cost to Admit LP call in n_{L}

In this section we show that for a given n_{H}, the cost to add an additional LP call in the system is monotonically non-decreasing and thus convex in n_{L}, for the LPAC states. Recall from Sec-

tion 3.4.3, we only need to look for an optimal LPAC policy whenever $K > A_{\mathrm{L}}$, therefore we prove Lemma 4 under this condition. Also, we only need to consider those LPAC states, i.e., $(n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_1$, such that, the states involved in the statement (3.21) of Lemma 4 belong to $\mathcal{S}$.

Lemma 4. *Let $K > A_{\mathrm{L}}$. $\forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_1$, s.t., $(n_{\mathrm{L}}+1, n_{\mathrm{H}}), (n_{\mathrm{L}}+2, n_{\mathrm{H}}) \in \mathcal{S}$, $\forall m \geq 0$:*

$$C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \leq C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}). \tag{3.21}$$

Proof. The proof of this lemma is given in Section 3.5.3. □

3.4.8 A Sufficient Condition for Inevitability of LP Preemption

Firstly we present Lemma 5, that provides an upper bound on the cost difference between two consecutive diagonal states, under certain conditions. This lemma helps us deduce Theorem 3 which states a sufficient condition for the inevitability of LP preemption, for the LPPC states. Here we assume that $\lambda_{\mathrm{L}} \leq \lambda_{\mathrm{H}}$. This situation normally occurs during an emergency situation, when call traffic of high priority talk groups, involved in life saving tasks, becomes higher than that of low priority talk groups, that are in a normal operating mode. Also note that it is only in Lemma 5, and Theorem 3 which is based on Lemma 5, that we assume $\lambda_{\mathrm{L}} \leq \lambda_{\mathrm{H}}$. The rest of the results in this study do not require this condition.

Lemma 5. *Let $\lambda_{\mathrm{L}} \leq \lambda_{\mathrm{H}}$ and $K \leq A_{\mathrm{L}} \leq A_{\mathrm{H}}$. $\forall m \geq 0$, $\forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}$, s.t., $\forall (n_{\mathrm{L}}-1, n_{\mathrm{H}}+1) \in \mathcal{S}$:*

$$C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \leq A_{\mathrm{H}} - K. \tag{3.22}$$

Proof. We prove this lemma by induction on m. The proof can be found in Section 3.5.4. □

Theorem 3. *Assume $\lambda_{\mathrm{L}} \leq \lambda_{\mathrm{H}}$ and $K \leq A_{\mathrm{L}} \leq A_{\mathrm{H}}$. Then, it is always optimal to admit an arriving HP call by preempting an LP call, whenever all channels are busy in the system with at least one LP call under service, i.e., $\forall m \geq 0$, $\forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_2$, if $K \leq A_{\mathrm{L}} \leq A_{\mathrm{H}}$ then:*

$$K + C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}}+1) \leq A_{\mathrm{H}} + C_m(n_{\mathrm{L}}, n_{\mathrm{H}}). \tag{3.23}$$

Proof. This follows directly from Lemma 5. □

Note that Lemma 5 is for all the states, but we also state it separately in this section, as Theorem 3 for LPPC states, because this helps us provide a sufficient condition for the inevitability of LP preemption.

3.4.9 A Particular Optimal LCAP Policy

Theorem 1 and 3 can be used together to derive a particular optimal LCAP policy, in the form of a corollary, as follows:

Corollary 1. *Assume* $\lambda_{\mathrm{L}} \leq \lambda_{\mathrm{H}}$ *and* $K \leq A_{\mathrm{L}} \leq A_{\mathrm{H}}$. *Then, the optimal LP call admission and preemption (LCAP) policy is as follows:*

1. **LPAC:** *Always admit an arriving LP call, whenever there is an idle channel in the system.*
2. **LPPC:** *Always admit an arriving HP call by preempting an LP call, whenever all channels are busy in the system with at least one LP call under service.*

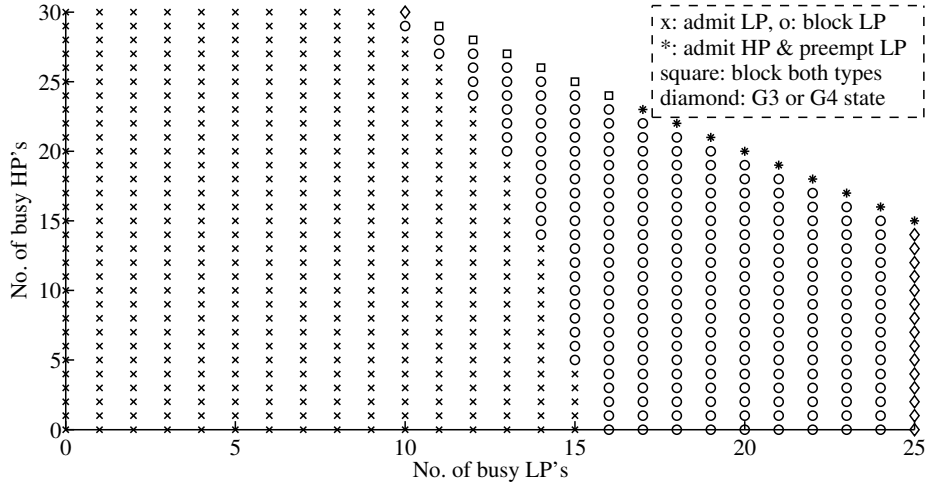
Proof. Part 1 of the corollary is implied by Theorem 1 and part 2 is implied by Theorem 3. □

Note that Corollary 1 provides sufficient conditions for a *myopic* policy [54] to be optimal.

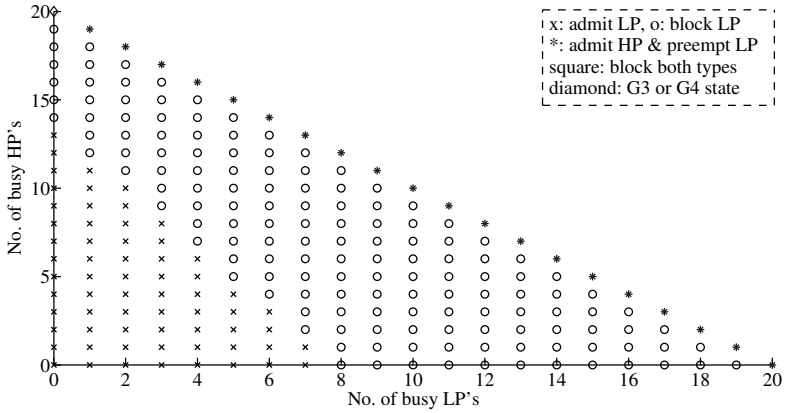
3.4.10 Numerically Searched Optimal Policy Results

In this section, we discuss the optimal policy search results for our system, presented in Figure 3.1, verifying our so far proved results. We present the results under the assumption $A_{\mathrm{L}} \leq A_{\mathrm{H}}$, due to the priority assignment of the calls, as mentioned in Section 3.2.2. We use the *policy iteration algorithm* [53], using the DP equations of our 2D Markov chain, in order to find the optimal LCAP decisions for each state. Also, our policy results correspond to $\alpha \to 0$ and the infinite horizon case. As described in detail in Section 3.2.2, though we have proved the results for the expected α-discounted cost over the finite horizon, our results also hold for $\alpha \to 0$ and the infinite horizon case.

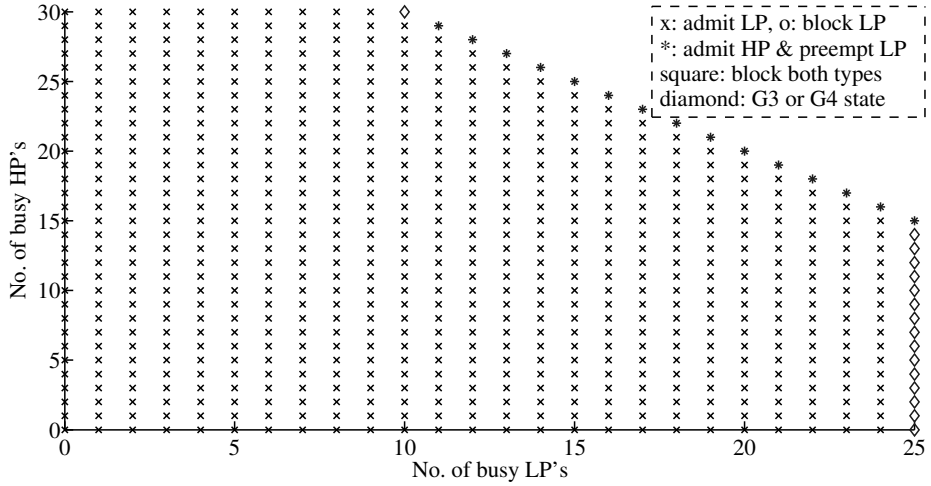
The results presented in Figure 3.1 show that the optimal LPAC policy is a state dependent threshold based policy, and thus verify Theorem 2. Tables 3.1 and 3.2 present the optimal state



(a) $N = 40$, Costs: $K = 8$, $A_{\mathrm{L}} = 2$, $A_{\mathrm{H}} = 4$.



(b) $N = 20$, Costs: $K = 6$, $A_{\mathrm{L}} = 4$, $A_{\mathrm{H}} = 5$



(c) $N = 40$, Costs: $K = 2$, $A_{\mathrm{L}} = 4$, $A_{\mathrm{H}} = 8$

FIGURE 3.1: Optimal policy search results ($M_{\mathrm{L}} = 25$, $M_{\mathrm{H}} = 30$, $\lambda_{\mathrm{L}} = 10$, $\lambda_{\mathrm{H}} = 12$, $\mu = 1$).

dependent LPAC thresholds for the cases presented in Figures 3.1(a) and 3.1(b) respectively. Since the thresholds depend on the number of LP calls in the system therefore these thresholds are termed as state dependent. Also, the LPAC policy in Figure 3.1(c) is that LP admission is optimal for all the LPAC states. This policy is a special case of a threshold based LPAC policy. For this particular case in Figure 3.1(c), $n_{\rm th}(n_{\rm L}) = 31, \forall n_{\rm L} \in \{0, 1, \ldots, 24\}$. Note that we do not include 25 in the set of all possible values of $n_{\rm L}$ while defining the LPAC threshold as $n_{\rm L} = 25$ is not included in set of LPAC states $\mathcal{S}_1$, for this case.

TABLE 3.1: Optimal LPAC Thresholds in Figure 3.1(a)

$n_{\rm L}$	$n_{\rm th}(n_{\rm L})$
$0 \leq n_{\rm L} \leq 9$	31
10	29
11	27
12	24
13	20
14	14
15	5
$16 \leq n_{\rm L} \leq 24$	0

TABLE 3.2: Optimal LPAC Thresholds in Figure 3.1(b)

$n_{\rm L}$	$n_{\rm th}(n_{\rm L})$
0	14
1	12
2	11
3	9
4	7
5	5
6	4
7	2
$8 \leq n_{\rm L} \leq 19$	0

Figure 3.1(c) clearly verifies Theorems 1 and 3, and Corollary 1. Note that it is only in Theorem 3, and Corollary 1 which is based on Theorem 3, that we assume $\lambda_{\rm L} \leq \lambda_{\rm H}$, as mentioned in Section 3.4.8. None of our other results requires this condition. Thus, even though rest of our

results hold for any values of arrival rates, for both classes, we only present numerical results for $\lambda_{\mathrm{L}} \leq \lambda_{\mathrm{H}}$ to avoid redundancy.

We observe in our numerically searched optimal policy results that the LPPC is a threshold based policy, as shown in Figure 3.1(a). Also, the LPPC policy in Figures 3.1(b) and 3.1(c) is a special case of a threshold based LPPC policy. This observation for the finite-source loss system is similar to [46] in which the authors demonstrate a similar result for the optimal preemption control, but for the Erlang loss system, using numerical examples without a proof. We can formally state our observation for the LPPC as follows:

Observation 1. *The optimal LPPC policy is a threshold policy, i.e., there exists an optimal threshold* $\hat{n}_{\mathrm{th}} \in \left\{0, 1, \ldots, \hat{M}_{\mathrm{H}}\right\}$*, such that, if an arriving HP call finds the system in state* $(n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_2$ *and* $n_{\mathrm{H}} < \hat{n}_{\mathrm{th}}$*, the HP call is admitted by preempting an LP call. Otherwise, the HP call is blocked.*

For the results presented in Figures 3.1(a), 3.1(b) and 3.1(c), the LPPC thresholds, i.e., the values of $\hat{n}_{\mathrm{th}}$, are 24, 20 and 30 respectively.

3.5 Proofs of Lemmas

3.5.1 Proof of Lemma 2

Proof. The methodology used for proving this lemma is also the *sample path argument*. Consider two systems, namely A and B. Assume that system A starts in state $(n_{\mathrm{L}}, n_{\mathrm{H}})$ and system B starts in state $(n_{\mathrm{L}}, n_{\mathrm{H}} + 1)$, with $(n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_1$. We couple the two systems such that all the call service and idle times, for the channels and sources, in both the systems are the same, except, the additional idle HP-source and channel in system A, and the additional busy HP-source and channel in system B. Also, assume that system A follows the optimal policy, while system B imitates all the decisions of system A, i.e., if system A admits (rejects) an LP call, system B also admits (rejects) it. Therefore, $C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) = C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}})$ and $C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) \leq C_m^{\mathrm{B}}(n_{\mathrm{L}}, n_{\mathrm{H}}+1)$, where $C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}})$ is the expected discounted cost of system A and $C_m^{\mathrm{B}}(n_{\mathrm{L}}, n_{\mathrm{H}}+1)$ is the expected discounted cost of system B, at time m. Now, we analyze all the possible cases, in the following

discussion, in which both the systems can completely couple, to become identical systems, and for each case we shall compare the costs of these systems.

1. The first case is, when both the systems end up with same costs after being completely coupled. For this case, there can only be two possible scenarios which enable systems A and B to couple completely. In the first scenario, at some point, the additional idle source of system A may generate an HP call and system A admits it to couple with system B. In the second scenario, at some point, the additional busy channel of system B completes the HP call service. In this way, the additional busy LP talk group and channel in system B become idle, and it completely couples with system A. For both these scenarios, the two systems couple completely without any change in their system costs. This is because, for our system model, we do not have any reward for call admission and completion that can cause the costs of both the systems to differ in the stated scenarios. Hence, for the first case:

$$C_m^{\mathrm{B}}(n_{\mathrm{L}}, n_{\mathrm{H}} + 1) = C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}}). \tag{3.24}$$

2. The second case is, when both the systems end up with different costs after being completely coupled. For this case, there can only be one possible scenario which enables systems A and B to couple completely. In this scenario, all the channels in system B are busy and an HP call arrives. Now, the only possibility that the two systems can couple completely, is when system B blocks it, provided the LP call preemption policy adopted by system B permits this, and system A admits the call without preemption, because system A still has an idle channel available. In this scenario, system B needs to pay an extra cost for blocking the HP call, i.e., A_{H}. This is the only difference in costs of both the systems. Therefore, for this case, we have:

$$C_m^{\mathrm{B}}(n_{\mathrm{L}}, n_{\mathrm{H}} + 1) - C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}}) = A_{\mathrm{H}}. \tag{3.25}$$

Based on our results in (3.24) and (3.25), we can now conclude:

$$C_m^{\mathrm{B}}(n_{\mathrm{L}}, n_{\mathrm{H}}+1) - C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}}) \leq A_{\mathrm{H}}. \tag{3.26}$$

According to our assumption, $C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) = C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}})$ and $C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) \leq C_m^{\mathrm{B}}(n_{\mathrm{L}}, n_{\mathrm{H}}+1)$. Therefore, we have:

$$C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \leq C_m^{\mathrm{B}}(n_{\mathrm{L}}, n_{\mathrm{H}}+1) - C_m^{\mathrm{A}}(n_{\mathrm{L}}, n_{\mathrm{H}}) \leq A_{\mathrm{H}}. \tag{3.27}$$

This completes the proof of Lemma 2. □

3.5.2 Proof of Lemma 3

Proof. This lemma can be proven by induction on m, the number of periods left in the horizon.

Step-1: Inequality (3.20) holds for $m = 0$, because, $C_0(n_{\mathrm{L}}, n_{\mathrm{H}}) = 0,\ \forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}$.

Induction step: Assume that for $m \geq 0$, inequality (3.20) holds.

Step-2: Assuming (3.20) holds for $m \geq 0$, we show that it holds for $m+1$ as well. There are three cases to consider. We make the distinction among these cases because each state in (3.20) may belong to a specific group of states, and each group has a different DP cost equation, as described in Section 3.3. We substitute the DP cost equations and use term by term comparison as follows.

Case-1: $(n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}_5$, s.t., $n_{\mathrm{L}} + n_{\mathrm{H}} \leq N - 3$. These are all the states in $\mathcal{S}_5$ excluding the diagonal states for which $n_{\mathrm{L}} + n_{\mathrm{H}} = N - 2$. The excluded diagonal states are analyzed in Cases 2 and 3. In this case, i.e., Case-1, all states in (3.20) are LPAC states, except when $M_{\mathrm{L}} < N$ and $n_{\mathrm{L}} = M_{\mathrm{L}} - 1$, which results in states $(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1)$ and $(n_{\mathrm{L}}, n_{\mathrm{H}}+1)$ as the no policy states. All these possibilities are catered by the following general expressions for this case.

$$\begin{aligned}
&C_{m+1}(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_{m+1}(n_{\mathrm{L}}, n_{\mathrm{H}}) \\
&= \Big([M_{\mathrm{L}} - (n_{\mathrm{L}}+1)]\, \lambda_{\mathrm{L}} \\
&\quad \times \Big[\min \{ C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) + A_{\mathrm{L}} \}
\end{aligned}$$

$$
\begin{aligned}
&- \min\left\{C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) + A_{\mathrm{L}}\right\}\Big]\Big) && (3.28)\\
&+ \Big(\left[M_{\mathrm{H}} - (n_{\mathrm{H}}+1)\right]\lambda_{\mathrm{H}} \times \left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1)\right]\Big) && (3.29)\\
&+ \lambda_{\mathrm{H}}\left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1)\right] && (3.30)\\
&+ n_{\mathrm{L}}\mu\left[C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}})\right] && (3.31)\\
&+ n_{\mathrm{H}}\mu\left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}-1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}-1)\right] && (3.32)\\
&+ \Big(\left[n_{\mathrm{L}}\lambda_{\mathrm{L}} + n_{\mathrm{H}}\lambda_{\mathrm{H}} + (N - n_{\mathrm{L}} - n_{\mathrm{H}} - 2)\,\mu\right] \times \left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}})\right]\Big) && (3.33)\\
&+ \mu\left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}})\right] && (3.34)\\
&+ \Big(\lambda_{\mathrm{L}}\Big[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - \min\left\{C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) + A_{\mathrm{L}}\right\}\Big]\Big) && (3.35)\\
\leq\ &\Big(\left[M_{\mathrm{L}} - (n_{\mathrm{L}}+1)\right]\lambda_{\mathrm{L}}\\
&\times \Big[\min\left\{C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}+1), C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) + A_{\mathrm{L}}\right\}\\
&- \min\left\{C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1), C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) + A_{\mathrm{L}}\right\}\Big]\Big) && (3.36)\\
&+ \Big(\left[M_{\mathrm{H}} - (n_{\mathrm{H}}+1)\right]\lambda_{\mathrm{H}} \times \left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+2) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+2)\right]\Big) && (3.37)\\
&+ \lambda_{\mathrm{H}}\left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1)\right] && (3.38)\\
&+ n_{\mathrm{L}}\mu\left[C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}}+1)\right] && (3.39)\\
&+ n_{\mathrm{H}}\mu\left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}})\right] && (3.40)\\
&+ \Big(\left[n_{\mathrm{L}}\lambda_{\mathrm{L}} + n_{\mathrm{H}}\lambda_{\mathrm{H}} + (N - n_{\mathrm{L}} - n_{\mathrm{H}} - 2)\,\mu\right]\\
&\times \left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1)\right]\Big) && (3.41)\\
&+ \mu\left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}})\right] && (3.42)\\
&+ \Big(\lambda_{\mathrm{L}}\Big[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1)\\
&- \min\left\{C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1), C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) + A_{\mathrm{L}}\right\}\Big]\Big) && (3.43)\\
=\ &C_{m+1}(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_{m+1}(n_{\mathrm{L}}, n_{\mathrm{H}}+1)
\end{aligned}
$$

Now we remark on the relations between the corresponding pairs of terms, on both sides of the inequality, to conclude the proof for Case-1. The relations (3.29) $\leq$ (3.37), (3.31) $\leq$ (3.39), (3.32) $\leq$ (3.40) and (3.33) $\leq$ (3.41) follow directly from the induction hypothesis (3.20). The relations (3.30) $=$ (3.38) and (3.34) $=$ (3.42) are obvious. However, the relations (3.28) $\leq$ (3.36) and (3.35) $\leq$ (3.43) need further justification, as detailed subsequently.

(3.28) $\leq$ (3.36):

$$\begin{aligned}
&\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}\}\\
&\quad-\min\{C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}),C_m(n_{\mathrm{L}},n_{\mathrm{H}})+A_{\mathrm{L}}\} &(3.44)\\
&=\big[\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}\}-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})\big]\\
&\quad-\big[\min\{C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}),C_m(n_{\mathrm{L}},n_{\mathrm{H}})+A_{\mathrm{L}}\}-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})\big] &(3.45)\\
&=\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}})-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}),A_{\mathrm{L}}\}\\
&\quad+\max\{0,C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})-C_m(n_{\mathrm{L}},n_{\mathrm{H}})-A_{\mathrm{L}}\} &(3.46)\\
&\overset{(3.20)}{\leq}\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}+1)-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1),A_{\mathrm{L}}\}\\
&\quad+\max\{0,C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1)-C_m(n_{\mathrm{L}},n_{\mathrm{H}}+1)-A_{\mathrm{L}}\}\\
&=\big[\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}+1),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1)+A_{\mathrm{L}}\}-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1)\big]\\
&\quad-\big[\min\{C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1),C_m(n_{\mathrm{L}},n_{\mathrm{H}}+1)+A_{\mathrm{L}}\}-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1)\big]\\
&=\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}+1),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1)+A_{\mathrm{L}}\}\\
&\quad-\min\{C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1),C_m(n_{\mathrm{L}},n_{\mathrm{H}}+1)+A_{\mathrm{L}}\}
\end{aligned}$$

We get (3.45) by adding and subtracting $C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})$ to (3.44). Then we convert *min* to *max*, to get (3.46). The same procedure is then applied in reverse to the other side of the inequality in order to justify that (3.28) $\leq$ (3.36) holds.

(3.35) $\leq$ (3.43):

$$\begin{aligned}
&C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})-\min\{C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}),C_m(n_{\mathrm{L}},n_{\mathrm{H}})+A_{\mathrm{L}}\}\\
&=\max\{0,C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})-C_m(n_{\mathrm{L}},n_{\mathrm{H}})-A_{\mathrm{L}}\}
\end{aligned}$$

$$\overset{(3.20)}{\leq} \max\{0, C_m(n_{\rm L}+1, n_{\rm H}+1) - C_m(n_{\rm L}, n_{\rm H}+1) - A_{\rm L}\}$$

$$= C_m(n_{\rm L}+1, n_{\rm H}+1) - \min\{C_m(n_{\rm L}+1, n_{\rm H}+1), C_m(n_{\rm L}, n_{\rm H}+1) + A_{\rm L}\}$$

This concludes Case-1.

Case-2: $M_{\rm H} < N, n_{\rm L} = N - M_{\rm H} - 1, n_{\rm H} = M_{\rm H} - 1$. This case deals with only one state. It is one of the $n_{\rm L}+n_{\rm H} = N-2$ diagonal states in $\mathcal{S}_5$. For example, state (9, 29) in Figures 3.1(a) and 3.1(c). The rest of the $n_{\rm L}+n_{\rm H} = N-2$ diagonal states in $\mathcal{S}_5$ are analyzed in Case-3. This case, i.e., Case-2, only exists when $M_{\rm H} < N$. Also, it is the only case that we need to consider separately. The rest of the two cases are applicable to all the possible scenarios under our assumption mentioned in Section 4.1, i.e., the total number of source-units is greater than the channels in the system, $M_{\rm L} + M_{\rm H} > N$. In this case, all the states in (3.20) are LPAC states, except $(n_{\rm L}+1, n_{\rm H}+1)$, which is a known-LPAC state.

$$\begin{aligned}
&C_{m+1}(n_{\rm L}+1, n_{\rm H}) - C_{m+1}(n_{\rm L}, n_{\rm H}) \\
&= \Big((M_{\rm L} + M_{\rm H} - N)\,\lambda_{\rm L} \\
&\quad \times \Big[\min\{C_m(n_{\rm L}+2, n_{\rm H}), C_m(n_{\rm L}+1, n_{\rm H}) + A_{\rm L}\} \\
&\quad - \min\{C_m(n_{\rm L}+1, n_{\rm H}), C_m(n_{\rm L}, n_{\rm H}) + A_{\rm L}\} \Big] \Big) && (3.47) \\
&\quad + \lambda_{\rm H}\,[C_m(n_{\rm L}+1, n_{\rm H}+1) - C_m(n_{\rm L}, n_{\rm H}+1)] && (3.48) \\
&\quad + (N - M_{\rm H} - 1)\,\mu\,[C_m(n_{\rm L}, n_{\rm H}) - C_m(n_{\rm L}-1, n_{\rm H})] && (3.49) \\
&\quad + (M_{\rm H} - 1)\,\mu\,[C_m(n_{\rm L}+1, n_{\rm H}-1) - C_m(n_{\rm L}, n_{\rm H}-1)] && (3.50) \\
&\quad + \Big([(N - M_{\rm H} - 1)\,\lambda_{\rm L} + (M_{\rm H} - 1)\,\lambda_{\rm H}] \times [C_m(n_{\rm L}+1, n_{\rm H}) - C_m(n_{\rm L}, n_{\rm H})] \Big) && (3.51) \\
&\quad + \mu\,[C_m(n_{\rm L}+1, n_{\rm H}) - C_m(n_{\rm L}, n_{\rm H})] && (3.52) \\
&\quad + \Big(\lambda_{\rm L} \Big[C_m(n_{\rm L}+1, n_{\rm H}) - \min\{C_m(n_{\rm L}+1, n_{\rm H}), C_m(n_{\rm L}, n_{\rm H}) + A_{\rm L}\} \Big] \Big) && (3.53) \\
&\leq \Big((M_{\rm L} + M_{\rm H} - N)\,\lambda_{\rm L} \Big[C_m(n_{\rm L}+1, n_{\rm H}+1) + A_{\rm L} \\
&\quad - \min\{C_m(n_{\rm L}+1, n_{\rm H}+1), C_m(n_{\rm L}, n_{\rm H}+1) + A_{\rm L}\} \Big] \Big) && (3.54)
\end{aligned}$$

$$+ \lambda_{\mathrm{H}} \left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) \right] \tag{3.55}$$

$$+ \Big((N - M_{\mathrm{H}} - 1)\, \mu \times \left[C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}}+1) \right] \Big) \tag{3.56}$$

$$+ (M_{\mathrm{H}} - 1)\, \mu \left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \right] \tag{3.57}$$

$$\begin{aligned} &+ \Big(\left[(N - M_{\mathrm{H}} - 1)\, \lambda_{\mathrm{L}} + (M_{\mathrm{H}} - 1)\, \lambda_{\mathrm{H}} \right] \\ &\times \left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) \right] \Big) \end{aligned} \tag{3.58}$$

$$+ \mu \left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \right] \tag{3.59}$$

$$\begin{aligned} &+ \Big(\lambda_{\mathrm{L}} \Big[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) \\ &- \min \left\{ C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1), C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) + A_{\mathrm{L}} \right\} \Big] \Big) \end{aligned} \tag{3.60}$$

$$= C_{m+1}(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_{m+1}(n_{\mathrm{L}}, n_{\mathrm{H}}+1)$$

The relations (3.49) $\leq$ (3.56), (3.50) $\leq$ (3.57) and (3.51) $\leq$ (3.58) follow directly from the induction hypothesis (3.20). The relations (3.48) $=$ (3.55) and (3.52) $=$ (3.59) are obvious. The relation (3.53) $\leq$ (3.60) is the same as (3.35) $\leq$ (3.43). However, the relation (3.47) $\leq$ (3.54) needs further justification.

(3.47) $\leq$ (3.54):

$$\begin{aligned} &\min \left\{ C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) + A_{\mathrm{L}} \right\} \\ &\quad - \min \left\{ C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) + A_{\mathrm{L}} \right\} \end{aligned} \tag{3.61}$$

$$\begin{aligned} &= \left[\min \left\{ C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) + A_{\mathrm{L}} \right\} - \left(C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) + A_{\mathrm{L}} \right) \right] \\ &\quad - \left[\min \left\{ C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) + A_{\mathrm{L}} \right\} - \left(C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) + A_{\mathrm{L}} \right) \right] \end{aligned} \tag{3.62}$$

$$\begin{aligned} &= \underbrace{\min \left\{ C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - A_{\mathrm{L}}, 0 \right\}}_{\leq 0} \\ &\quad + \max \left\{ A_{\mathrm{L}}, C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \right\} \end{aligned} \tag{3.63}$$

$$\begin{aligned} &\overset{(3.20)}{\leq} 0 + \max \left\{ A_{\mathrm{L}}, C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) \right\} \\ &= C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) + A_{\mathrm{L}} - \min \left\{ C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1), C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) + A_{\mathrm{L}} \right\} \end{aligned}$$

We get (3.62) by adding and subtracting $C_m(n_{\rm L}+1, n_{\rm H}) + A_{\rm L}$ to (3.61). Then we convert *min* to *max*, to get (3.63). The same procedure is then applied in reverse to the other side of the inequality in order to justify that (3.47) $\leq$ (3.54) holds. This concludes Case-2.

Case-3: $(n_{\rm L}, n_{\rm H}) \in \mathcal{S}_5$, s.t., $N - \hat{M}_{\rm H} \leq n_{\rm L} \leq \hat{M}_{\rm L} - 1$, $n_{\rm L} + n_{\rm H} = N - 2$. These are the $n_{\rm L} + n_{\rm H} = N - 2$ diagonal states in $\mathcal{S}_5$ excluding the Case-2 state. Note that the Case-2 state only exists when $M_{\rm H} < N$. In this case, i.e. Case-3, states $(n_{\rm L}, n_{\rm H})$ and $(n_{\rm L}, n_{\rm H}+1)$, in (3.20), are LPAC states, and $(n_{\rm L}+1, n_{\rm H}+1)$ is an LPPC state. However, $(n_{\rm L}+1, n_{\rm H})$ is an LPAC state, except when $M_{\rm L} < N$ and $n_{\rm L} = M_{\rm L} - 1$, which results in state $(n_{\rm L}+1, n_{\rm H})$ as a no policy state. All these possibilities are catered by the following general expressions for this case.

$$
\begin{aligned}
&C_{m+1}(n_{\rm L}+1, n_{\rm H}) - C_{m+1}(n_{\rm L}, n_{\rm H}) \\
&= \Big(\left[M_{\rm L} - (n_{\rm L}+1)\right] \lambda_{\rm L} \\
&\quad \times \Big[\min\left\{C_m(n_{\rm L}+2, n_{\rm H}), C_m(n_{\rm L}+1, n_{\rm H}) + A_{\rm L}\right\} \\
&\quad - \min\left\{C_m(n_{\rm L}+1, n_{\rm H}), C_m(n_{\rm L}, n_{\rm H}) + A_{\rm L}\right\} \Big] \Big) && (3.64) \\
&\quad + \Big(\left[M_{\rm H} - (n_{\rm H}+1)\right] \lambda_{\rm H} \times \left[C_m(n_{\rm L}+1, n_{\rm H}+1) - C_m(n_{\rm L}, n_{\rm H}+1)\right] \Big) && (3.65) \\
&\quad + \lambda_{\rm H} \left[C_m(n_{\rm L}+1, n_{\rm H}+1) - C_m(n_{\rm L}, n_{\rm H}+1)\right] && (3.66) \\
&\quad + n_{\rm L}\mu \left[C_m(n_{\rm L}, n_{\rm H}) - C_m(n_{\rm L}-1, n_{\rm H})\right] && (3.67) \\
&\quad + n_{\rm H}\mu \left[C_m(n_{\rm L}+1, n_{\rm H}-1) - C_m(n_{\rm L}, n_{\rm H}-1)\right] && (3.68) \\
&\quad + \left[n_{\rm L}\lambda_{\rm L} + n_{\rm H}\lambda_{\rm H}\right] \left[C_m(n_{\rm L}+1, n_{\rm H}) - C_m(n_{\rm L}, n_{\rm H})\right] && (3.69) \\
&\quad + \mu \left[C_m(n_{\rm L}+1, n_{\rm H}) - C_m(n_{\rm L}, n_{\rm H})\right] && (3.70) \\
&\quad + \Big(\lambda_{\rm L} \Big[C_m(n_{\rm L}+1, n_{\rm H}) - \min\left\{C_m(n_{\rm L}+1, n_{\rm H}), C_m(n_{\rm L}, n_{\rm H}) + A_{\rm L}\right\} \Big] \Big) && (3.71) \\
&\leq \Big(\left[M_{\rm L} - (n_{\rm L}+1)\right] \lambda_{\rm L} \\
&\quad \times \Big[C_m(n_{\rm L}+1, n_{\rm H}+1) + A_{\rm L} \\
&\quad - \min\left\{C_m(n_{\rm L}+1, n_{\rm H}+1), C_m(n_{\rm L}, n_{\rm H}+1) + A_{\rm L}\right\} \Big] \Big) && (3.72)
\end{aligned}
$$

$$+\Bigg([M_{\rm H}-(n_{\rm H}+1)]\,\lambda_{\rm H}$$

$$\times\Big[\min\left\{C_m(n_{\rm L}+1,n_{\rm H}+1)+A_{\rm H},C_m(n_{\rm L},n_{\rm H}+2)+K\right\}-C_m(n_{\rm L},n_{\rm H}+2)\Big]\Bigg) \quad (3.73)$$

$$+\lambda_{\rm H}\left[C_m(n_{\rm L}+1,n_{\rm H}+1)-C_m(n_{\rm L},n_{\rm H}+1)\right] \quad (3.74)$$

$$+n_{\rm L}\mu\left[C_m(n_{\rm L},n_{\rm H}+1)-C_m(n_{\rm L}-1,n_{\rm H}+1)\right] \quad (3.75)$$

$$+n_{\rm H}\mu\left[C_m(n_{\rm L}+1,n_{\rm H})-C_m(n_{\rm L},n_{\rm H})\right] \quad (3.76)$$

$$+\Big([n_{\rm L}\lambda_{\rm L}+n_{\rm H}\lambda_{\rm H}]\times[C_m(n_{\rm L}+1,n_{\rm H}+1)-C_m(n_{\rm L},n_{\rm H}+1)]\Big) \quad (3.77)$$

$$+\mu\left[C_m(n_{\rm L}+1,n_{\rm H})-C_m(n_{\rm L},n_{\rm H})\right] \quad (3.78)$$

$$+\Big(\lambda_{\rm L}\Big[C_m(n_{\rm L}+1,n_{\rm H}+1)$$

$$-\min\left\{C_m(n_{\rm L}+1,n_{\rm H}+1),C_m(n_{\rm L},n_{\rm H}+1)+A_{\rm L}\right\}\Big]\Big) \quad (3.79)$$

$$=C_{m+1}(n_{\rm L}+1,n_{\rm H}+1)-C_{m+1}(n_{\rm L},n_{\rm H}+1)$$

The relations , (3.67) $\leq$ (3.75), (3.68) $\leq$ (3.76) and (3.69) $\leq$ (3.77) follow directly from the induction hypothesis (3.20). The relations (3.66) = (3.74) and (3.70) = (3.78) are obvious. Justifications of relations (3.64) $\leq$ (3.72) and (3.71) $\leq$ (3.79) are same as that of (3.47) $\leq$ (3.54) and (3.35) $\leq$ (3.43), respectively. However, (3.65) $\leq$ (3.73) needs further justification.

(3.73) $\geq$ (3.65): There are only two possible cases. For the first one, we assume that the LP call preemption is the optimal LPPC policy in state $(n_{\rm L}+1,n_{\rm H}+1)$. In this case:

$$\begin{aligned}
&\min\left\{C_m(n_{\rm L}+1,n_{\rm H}+1)+A_{\rm H},C_m(n_{\rm L},n_{\rm H}+2)+K\right\}-C_m(n_{\rm L},n_{\rm H}+2)\\
&=[C_m(n_{\rm L},n_{\rm H}+2)+K]-C_m(n_{\rm L},n_{\rm H}+2)\\
&=K\\
&\geq C_m(n_{\rm L}+1,n_{\rm H}+1)-C_m(n_{\rm L},n_{\rm H}+1). \qquad (3.80)
\end{aligned}$$

Here, (3.80) results from Lemma 1.2 after replacing $(n_{\rm L},n_{\rm H})$ with $(n_{\rm L},n_{\rm H}+1)$ in the lemma, under our assumption for Lemma 3, i.e, $K>A_{\rm L}$.

Now, for the other case, we assume that the LP call preemption is not the optimal LPPC policy in state $(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1)$. In this case:

$$\begin{aligned}
&\min\left\{C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1)+A_{\mathrm{H}}, C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+2)+K\right\} - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+2) \\
&= \left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1)+A_{\mathrm{H}}\right] - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+2) \qquad (3.81) \\
&\geq C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1). \qquad (3.82)
\end{aligned}$$

Here, (3.82) is obtained by using inequality (3.18) of Lemma 2 for A_{H} in (3.81), after replacing $(n_{\mathrm{L}}, n_{\mathrm{H}})$ with $(n_{\mathrm{L}}, n_{\mathrm{H}}+1)$ in (3.18). This concludes the proof for Lemma 3. □

3.5.3 Proof of Lemma 4

Proof. This lemma can be proven by induction on m, the number of periods left in the horizon.

Step-1: Inequality (3.21) holds for $m=0$, because, $C_0(n_{\mathrm{L}}, n_{\mathrm{H}}) = 0,\ \forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}$.

Induction step: Assume that for $m \geq 0$, inequality (3.21) holds.

Step-2: Assuming (3.21) holds for $m \geq 0$, we show that it holds for $m+1$ as well. There are two cases to consider. We make the distinction among these cases because each state in (3.21) may belong to a specific group of states, and each group has a different DP cost equation, as described in Section 3.3. We substitute the DP equations and use term by term comparison as shown in the following cases.

Case-1: Let $n_{\mathrm{L}} + n_{\mathrm{H}} \leq N-3$. In this case, i.e., Case-1, all states in (3.21) are LPAC states, except when $M_{\mathrm{L}} < N$ and $n_{\mathrm{L}} = M_{\mathrm{L}}$, which results in state $(n_{\mathrm{L}}+2, n_{\mathrm{H}})$ as a no policy state. All these possibilities are catered by the following general expressions for this case.

$$\begin{aligned}
&C_{m+1}(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_{m+1}(n_{\mathrm{L}}, n_{\mathrm{H}}) \\
&= \Big(\left[M_{\mathrm{L}} - (n_{\mathrm{L}}+2)\right] \lambda_{\mathrm{L}} \\
&\quad \times \Big[\min\left\{C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) + A_{\mathrm{L}}\right\} \\
&\quad - \min\left\{C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) + A_{\mathrm{L}}\right\} \Big] \Big) \qquad (3.83) \\
&\quad + \left(M_{\mathrm{H}} - n_{\mathrm{H}}\right) \lambda_{\mathrm{H}} \times \left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1)\right] \qquad (3.84)
\end{aligned}$$

$$+ n_{\mathrm{L}}\mu\left[C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}})\right] \tag{3.85}$$

$$+ n_{\mathrm{H}}\mu\left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}-1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}-1)\right] \tag{3.86}$$

$$+ \Big(\left[n_{\mathrm{L}}\lambda_{\mathrm{L}} + n_{\mathrm{H}}\lambda_{\mathrm{H}} + (N - n_{\mathrm{L}} - n_{\mathrm{H}} - 2)\,\mu\right] \times \left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}})\right] \Big) \tag{3.87}$$

$$+ \mu\left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}})\right] \tag{3.88}$$

$$+ \Bigg(\lambda_{\mathrm{L}} \Bigg[\min\left\{C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) + A_{\mathrm{L}}\right\}$$

$$- 2 \times \min\left\{C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) + A_{\mathrm{L}}\right\} + C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) \Bigg] \Bigg) \tag{3.89}$$

$$\leq \Big(\left[M_{\mathrm{L}} - (n_{\mathrm{L}}+2)\right] \lambda_{\mathrm{L}}$$

$$\times \Big[\min\left\{C_m(n_{\mathrm{L}}+3, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}) + A_{\mathrm{L}}\right\}$$

$$- \min\left\{C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) + A_{\mathrm{L}}\right\} \Big] \Big) \tag{3.90}$$

$$+ \left(M_{\mathrm{H}} - n_{\mathrm{H}}\right) \lambda_{\mathrm{H}} \times \left[C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1)\right] \tag{3.91}$$

$$+ n_{\mathrm{L}}\mu\left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}})\right] \tag{3.92}$$

$$+ n_{\mathrm{H}}\mu\left[C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}-1) - C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}-1)\right] \tag{3.93}$$

$$+ \Big(\left[n_{\mathrm{L}}\lambda_{\mathrm{L}} + n_{\mathrm{H}}\lambda_{\mathrm{H}} + (N - n_{\mathrm{L}} - n_{\mathrm{H}} - 2)\,\mu\right]$$

$$\times \left[C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}})\right] \Big) \tag{3.94}$$

$$+ \mu\left[C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}})\right] \tag{3.95}$$

$$+ \Bigg(\lambda_{\mathrm{L}} \Bigg[- \min\left\{C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) + A_{\mathrm{L}}\right\}$$

$$+ 2\,C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}) \Bigg] \Bigg) \tag{3.96}$$

$$= C_{m+1}(n_{\mathrm{L}}+2, n_{\mathrm{H}}) - C_{m+1}(n_{\mathrm{L}}+1, n_{\mathrm{H}})$$

Now we remark on the relations between the corresponding pairs of terms, on both sides of the inequality, to conclude the proof for Case-1. The relations (3.84) $\leq$ (3.91), (3.85) $\leq$ (3.92), (3.86) $\leq$ (3.93) and (3.87) $\leq$ (3.94) follow directly from the induction hypothesis (3.21). The

relation (3.88) = (3.95) is obvious. However, the relations (3.83) $\leq$ (3.90) and (3.89) $\leq$ (3.96) need further justification, as detailed subsequently.

(3.83) $\leq$ (3.90):

$$\begin{aligned}
&\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}\}\\
&\quad-\min\{C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}),C_m(n_{\mathrm{L}},n_{\mathrm{H}})+A_{\mathrm{L}}\} &(3.97)\\
&=\big[\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}\}-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})\big]\\
&\quad-\big[\min\{C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}),C_m(n_{\mathrm{L}},n_{\mathrm{H}})+A_{\mathrm{L}}\}-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})\big] &(3.98)\\
&=\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}})-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}),A_{\mathrm{L}}\}\\
&\quad+\max\{0,C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})-C_m(n_{\mathrm{L}},n_{\mathrm{H}})-A_{\mathrm{L}}\} &(3.99)\\
&\overset{(3.21)}{\leq}\min\{C_m(n_{\mathrm{L}}+3,n_{\mathrm{H}})-C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),A_{\mathrm{L}}\}\\
&\quad+\max\{0,C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}})-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})-A_{\mathrm{L}}\}\\
&=\big[\min\{C_m(n_{\mathrm{L}}+3,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}})+A_{\mathrm{L}}\}-C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}})\big]\\
&\quad-\big[\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}\}-C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}})\big]\\
&=\min\{C_m(n_{\mathrm{L}}+3,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}})+A_{\mathrm{L}}\}\\
&\quad-\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}\}
\end{aligned}$$

We get (3.98) by adding and subtracting $C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})$ to (3.97). Then we convert *min* to *max*, to get (3.99). The same procedure is then applied in reverse to the other side of the inequality in order to justify that (3.83) $\leq$ (3.90) holds.

(3.89) $\leq$ (3.96):

$$\begin{aligned}
&\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}\}\\
&\quad-2\times\min\{C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}),C_m(n_{\mathrm{L}},n_{\mathrm{H}})+A_{\mathrm{L}}\}+C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}) &(3.100)\\
&=\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}\}-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})\\
&\quad+2\Big[-\min\{C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}),C_m(n_{\mathrm{L}},n_{\mathrm{H}})+A_{\mathrm{L}}\}+C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})\Big] &(3.101)
\end{aligned}$$

$$= \min\{C_m(n_{\rm L}+2,n_{\rm H}), C_m(n_{\rm L}+1,n_{\rm H})+A_{\rm L}\} - C_m(n_{\rm L}+1,n_{\rm H}) \tag{3.102}$$

$$+ 2\Big[\max\{0, C_m(n_{\rm L}+1,n_{\rm H}) - C_m(n_{\rm L},n_{\rm H}) - A_{\rm L}\}\Big] \tag{3.103}$$

$$\overset{(3.21)}{\leq} \min\{C_m(n_{\rm L}+2,n_{\rm H}), C_m(n_{\rm L}+1,n_{\rm H})+A_{\rm L}\} - C_m(n_{\rm L}+1,n_{\rm H}) \tag{3.104}$$

$$+ 2\Big[\max\{0, C_m(n_{\rm L}+2,n_{\rm H}) - C_m(n_{\rm L}+1,n_{\rm H}) - A_{\rm L}\}\Big] \tag{3.105}$$

$$= \min\{C_m(n_{\rm L}+2,n_{\rm H}), C_m(n_{\rm L}+1,n_{\rm H})+A_{\rm L}\} - C_m(n_{\rm L}+1,n_{\rm H})$$
$$+ 2\Big[-\min\{C_m(n_{\rm L}+2,n_{\rm H}), C_m(n_{\rm L}+1,n_{\rm H})+A_{\rm L}\} + C_m(n_{\rm L}+2,n_{\rm H})\Big] \tag{3.106}$$

$$= -\min\{C_m(n_{\rm L}+2,n_{\rm H}), C_m(n_{\rm L}+1,n_{\rm H})+A_{\rm L}\}$$
$$+ 2\,C_m(n_{\rm L}+2,n_{\rm H}) - C_m(n_{\rm L}+1,n_{\rm H})$$

We get (3.101) by adding and subtracting $C_m(n_{\rm L}+1,n_{\rm H})$ to (3.100). Then we convert *min* to *max*, to get (3.103). Note that (3.102) = (3.104) and according to the induction hypothesis (3.103) $\leq$ (3.105). We then convert *max* to *min*, and also add and subtract $C_m(n_{\rm L}+2,n_{\rm H})$, to get (3.106). This concludes Case-1.

Case-2: Let $n_{\rm L}+n_{\rm H} = N-2$. In this case, i.e., Case-2, all states in (3.21) are LPAC states, except $(n_{\rm L}+2,n_{\rm H})$, which can either be an LPPC or a known-LPAC state. All these possibilities are catered by the following general expressions for this case.

$$C_{m+1}(n_{\rm L}+1,n_{\rm H}) - C_{m+1}(n_{\rm L},n_{\rm H})$$
$$= \Big([M_{\rm L}-(n_{\rm L}+2)]\,\lambda_{\rm L}$$
$$\times\Big[\min\{C_m(n_{\rm L}+2,n_{\rm H}), C_m(n_{\rm L}+1,n_{\rm H})+A_{\rm L}\}$$
$$-\min\{C_m(n_{\rm L}+1,n_{\rm H}), C_m(n_{\rm L},n_{\rm H})+A_{\rm L}\}\Big]\Big) \tag{3.107}$$

$$+ (M_{\rm H}-n_{\rm H})\,\lambda_{\rm H} \times [C_m(n_{\rm L}+1,n_{\rm H}+1) - C_m(n_{\rm L},n_{\rm H}+1)] \tag{3.108}$$

$$+ n_{\rm L}\mu\,[C_m(n_{\rm L},n_{\rm H}) - C_m(n_{\rm L}-1,n_{\rm H})] \tag{3.109}$$

$$+ n_{\rm H}\mu\,[C_m(n_{\rm L}+1,n_{\rm H}-1) - C_m(n_{\rm L},n_{\rm H}-1)] \tag{3.110}$$

$$+ \Big([n_{\rm L}\lambda_{\rm L} + n_{\rm H}\lambda_{\rm H}] \times [C_m(n_{\rm L}+1,n_{\rm H}) - C_m(n_{\rm L},n_{\rm H})]\Big) \tag{3.111}$$

$$+ \mu \left[C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \right] \tag{3.112}$$

$$+ \Bigg(\lambda_{\mathrm{L}} \Bigg[\min \left\{ C_m(n_{\mathrm{L}} + 2, n_{\mathrm{H}}), C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}}) + A_{\mathrm{L}} \right\}$$

$$- 2 \times \min \left\{ C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}}), C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) + A_{\mathrm{L}} \right\} + C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}}) \Bigg] \Bigg) \tag{3.113}$$

$$\leq \Big(\left[M_{\mathrm{L}} - (n_{\mathrm{L}} + 2) \right] \lambda_{\mathrm{L}}$$

$$\times \Big[C_m(n_{\mathrm{L}} + 2, n_{\mathrm{H}}) + A_{\mathrm{L}} - \min \left\{ C_m(n_{\mathrm{L}} + 2, n_{\mathrm{H}}), C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}}) + A_{\mathrm{L}} \right\} \Big] \Big) \tag{3.114}$$

$$+ \Bigg(\left(M_{\mathrm{H}} - n_{\mathrm{H}} \right) \lambda_{\mathrm{H}} \Bigg[\min \left\{ C_m(n_{\mathrm{L}} + 2, n_{\mathrm{H}}) + A_{\mathrm{H}}, C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}} + 1) + K \right\}$$

$$- C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}} + 1) \Bigg] \Bigg) \tag{3.115}$$

$$+ n_{\mathrm{L}} \mu \left[C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \right] \tag{3.116}$$

$$+ n_{\mathrm{H}} \mu \left[C_m(n_{\mathrm{L}} + 2, n_{\mathrm{H}} - 1) - C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}} - 1) \right] \tag{3.117}$$

$$+ \Big(\left[n_{\mathrm{L}} \lambda_{\mathrm{L}} + n_{\mathrm{H}} \lambda_{\mathrm{H}} \right]$$

$$\times \left[C_m(n_{\mathrm{L}} + 2, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}}) \right] \Big) \tag{3.118}$$

$$+ \mu \left[C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}) \right] \tag{3.119}$$

$$+ \Bigg(\lambda_{\mathrm{L}} \Bigg[- \min \left\{ C_m(n_{\mathrm{L}} + 2, n_{\mathrm{H}}), C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}}) + A_{\mathrm{L}} \right\}$$

$$+ 2\, C_m(n_{\mathrm{L}} + 2, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}} + 1, n_{\mathrm{H}}) \Bigg] \Bigg) \tag{3.120}$$

$$= C_{m+1}(n_{\mathrm{L}} + 2, n_{\mathrm{H}}) - C_{m+1}(n_{\mathrm{L}} + 1, n_{\mathrm{H}})$$

The relations (3.109) $\leq$ (3.116), (3.110) $\leq$ (3.117) and (3.111) $\leq$ (3.118) follow directly from the induction hypothesis (3.21). The relation (3.113) $\leq$ (3.120) is same as (3.89) $\leq$ (3.96), and the relation (3.112) $=$ (3.119) is obvious. However, the relations (3.107) $\leq$ (3.114) and (3.108) $\leq$ (3.115) need further justification. In the subsequent discussion, we provide details for this justification.

(3.107) $\leq$ (3.114): This result is proved as follows.

$$\begin{aligned}
&\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}\}\\
&\quad-\min\{C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}),C_m(n_{\mathrm{L}},n_{\mathrm{H}})+A_{\mathrm{L}}\} &(3.121)\\
&=\big[\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}\}-(C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}})\big]\\
&\quad-\big[\min\{C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}),C_m(n_{\mathrm{L}},n_{\mathrm{H}})+A_{\mathrm{L}}\}-(C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}})\big] &(3.122)\\
&=\underbrace{\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}})-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})-A_{\mathrm{L}},0\}}_{\leq 0}\\
&\quad+\max\{A_{\mathrm{L}},C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})-C_m(n_{\mathrm{L}},n_{\mathrm{H}})\} &(3.123)\\
&\overset{(3.21)}{\leq}\ 0+\max\{A_{\mathrm{L}},C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}})-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})\}\\
&=C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}})+A_{\mathrm{L}}-\min\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}}),C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}\}
\end{aligned}$$

We get (3.122) by adding and subtracting $C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}})+A_{\mathrm{L}}$ to (3.121). Then we convert *min* to *max*, to get (3.123). The same procedure is then applied in reverse to the other side of the inequality in order to justify that (3.107) $\leq$ (3.114) holds.

(3.115) $\geq$ (3.108): There are only two possible cases. For the first one, we assume that the LP call preemption is the optimal LPPC policy in state $(n_{\mathrm{L}}+2,n_{\mathrm{H}})$. In this case:

$$\begin{aligned}
&\min\big\{C_m(n_{\mathrm{L}}+2,n_{\mathrm{H}})+A_{\mathrm{H}},C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1)+K\big\}-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1)\\
&=[C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1)+K]-C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1)\\
&=K\\
&\geq C_m(n_{\mathrm{L}}+1,n_{\mathrm{H}}+1)-C_m(n_{\mathrm{L}},n_{\mathrm{H}}+1). &(3.124)
\end{aligned}$$

Here, (3.124) results from Lemma 1.2 after replacing $(n_{\mathrm{L}},n_{\mathrm{H}})$ with $(n_{\mathrm{L}},n_{\mathrm{H}}+1)$ in the lemma, under our assumption for Lemma 4, i.e, $K>A_{\mathrm{L}}$.

Now, for the other case, we assume that the LP call preemption is not the optimal LPPC policy in state $(n_{\mathrm{L}}+2,n_{\mathrm{H}})$. In this case, (3.115) $\geq$ (3.108) can be proved as demonstrated by the subsequent mathematical expressions. Note that, (3.126) is obtained by using inequality (3.18)

of Lemma 2 for A_{H} in (3.125), after replacing $(n_{\mathrm{L}}, n_{\mathrm{H}})$ with $(n_{\mathrm{L}}+2, n_{\mathrm{H}})$ in (3.18). Finally we get (3.127) using the induction hypothesis.

$$\min\left\{C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}) + A_{\mathrm{H}}, C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) + K\right\} - C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1)$$

$$= [C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}) + A_{\mathrm{H}}] - C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) \tag{3.125}$$

$$\geq C_m(n_{\mathrm{L}}+2, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1). \tag{3.126}$$

$$\overset{(3.21)}{\geq} C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1). \tag{3.127}$$

This concludes the proof for Lemma 4. □

3.5.4 Proof of Lemma 5

Proof. This lemma can be proven by induction on m, the number of periods left in the horizon.

Step-1: Inequality (3.22) holds for $m = 0$, because, $C_0(n_{\mathrm{L}}, n_{\mathrm{H}}) = 0,\ \forall (n_{\mathrm{L}}, n_{\mathrm{H}}) \in \mathcal{S}$.

Induction step: Assume that for $m \geq 0$, inequality (3.22) holds. This also implies that it is optimal to admit an HP call by preempting an LP call for all LPPC states, under stated conditions. The statement of Theorem 3 can help understand this in a better way.

Step-2: Assuming (3.22) holds for $m \geq 0$, we show that it holds for $m+1$ as well. There are two cases to consider. We make the distinction among these cases because each state in (3.22) may belong to a specific group of states, and each group has a different DP cost equation, as described in Section 3.3. We substitute the DP equations and use term by term comparison as shown in the following cases. Note that due to the assumption $K \leq A_{\mathrm{L}}$, it is optimal to admit an LP call for all LPAC states, according to Theorem 1. Also recall that the induction hypothesis implies that it is optimal to admit an HP call by preempting an LP call for all LPPC states, under stated conditions. We write all the DP cost equations in the following cases according to these optimal policy actions.

Case-1: Let $n_{\mathrm{L}} + n_{\mathrm{H}} \leq N - 1$. In this case, i.e., Case-1, all states in (3.22) are LPAC states, except when $M_{\mathrm{L}} < N$ and $n_{\mathrm{L}} = M_{\mathrm{L}}$, which results in state $(n_{\mathrm{L}}, n_{\mathrm{H}})$ as a no policy state. All these possibilities are catered by the following general expressions for this case. Note that

here we use the induction hypothesis to apply the upper bound for all the terms involving the cost difference between consecutive diagonal states. We also use the uniformization assumption $M_{\mathrm{L}}\lambda_L + M_{\mathrm{H}}\lambda_{\mathrm{H}} + N\mu + \alpha = 1$, along with the assumption that $\lambda_{\mathrm{L}} \leq \lambda_{\mathrm{H}}$.

$$
\begin{aligned}
&C_{m+1}(n_{\mathrm{L}}-1, n_{\mathrm{H}}+1) - C_{m+1}(n_{\mathrm{L}}, n_{\mathrm{H}}) \\
&= (M_{\mathrm{L}} - n_{\mathrm{L}})\lambda_{\mathrm{L}} \left[C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}+1, n_{\mathrm{H}})\right] \\
&+ \left[M_{\mathrm{H}} - (n_{\mathrm{H}}+1)\right] \lambda_{\mathrm{H}} \left[C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}}+2) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1)\right] \\
&+ (n_{\mathrm{L}}-1)\mu \left[C_m(n_{\mathrm{L}}-2, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}})\right] \\
&+ n_{\mathrm{H}}\mu \left[C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}}) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}}-1)\right] \\
&+ \left[n_{\mathrm{L}}\lambda_{\mathrm{L}} + n_{\mathrm{H}}\lambda_{\mathrm{H}} + (N - n_{\mathrm{L}} - n_{\mathrm{H}})\mu\right] \times \left[C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}})\right] \\
&+ \underbrace{\lambda_{\mathrm{L}} \left[C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}}+1)\right] - \lambda_{\mathrm{H}} \left[C_m(n_{\mathrm{L}}, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}}+1)\right]}_{\leq 0, \text{ as } \lambda_{\mathrm{L}} \leq \lambda_H}
\end{aligned}
\tag{3.128}
$$

$$
\overset{(3.22)}{\leq} \underbrace{\left[M_{\mathrm{L}}\lambda_L + (M_{\mathrm{H}}-1)\lambda_{\mathrm{H}} - \mu + N\mu\right]}_{=1-\alpha-\lambda_{\mathrm{H}}-\mu, \text{ as } M_{\mathrm{L}}\lambda_L + M_{\mathrm{H}}\lambda_{\mathrm{H}} + N\mu + \alpha = 1} \times (A_{\mathrm{H}} - K) + 0 \tag{3.129}
$$

$$
= \underbrace{(1 - \alpha - \lambda_{\mathrm{H}} - \mu)}_{\leq 1, \text{ as } M_{\mathrm{L}}\lambda_L + M_{\mathrm{H}}\lambda_{\mathrm{H}} + N\mu + \alpha = 1} \times (A_{\mathrm{H}} - K) \tag{3.130}
$$

$$
\leq A_{\mathrm{H}} - K. \tag{3.131}
$$

This concludes Case-1.

Case-2: Let $n_{\mathrm{L}} + n_{\mathrm{H}} = N$. In this case, i.e., Case-2, all states in (3.22) are LPPC states, except when $M_{\mathrm{H}} < N$ and $n_{\mathrm{H}} = M_{\mathrm{H}} - 1$, which results in state $(n_{\mathrm{L}} - 1, n_{\mathrm{H}} + 1)$ as the known-LPAC state. All these possibilities are catered by the following general expressions for this case.

$$
\begin{aligned}
&C_{m+1}(n_{\mathrm{L}}-1, n_{\mathrm{H}}+1) - C_{m+1}(n_{\mathrm{L}}, n_{\mathrm{H}}) \\
&= (M_{\mathrm{L}} - n_{\mathrm{L}})\lambda_{\mathrm{L}} \left[C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}, n_{\mathrm{H}})\right] \\
&+ \left[M_{\mathrm{H}} - (n_{\mathrm{H}}+1)\right] \lambda_{\mathrm{H}} \left[C_m(n_{\mathrm{L}}-2, n_{\mathrm{H}}+2) - C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}}+1)\right] \\
&+ (n_{\mathrm{L}}-1)\mu \left[C_m(n_{\mathrm{L}}-2, n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}}-1, n_{\mathrm{H}})\right]
\end{aligned}
$$

$$+ n_{\mathrm{H}}\mu\left[C_m(n_{\mathrm{L}}-1,n_{\mathrm{H}}) - C_m(n_{\mathrm{L}},n_{\mathrm{H}}-1)\right]$$

$$+ \left[n_{\mathrm{L}}\lambda_{\mathrm{L}} + n_{\mathrm{H}}\lambda_{\mathrm{H}}\right] \times \left[C_m(n_{\mathrm{L}}-1,n_{\mathrm{H}}+1) - C_m(n_{\mathrm{L}},n_{\mathrm{H}})\right]$$

$$+ \underbrace{\lambda_{\mathrm{L}}A_{\mathrm{L}}}_{\leq \lambda_{\mathrm{H}}A_{\mathrm{H}},\text{ as } \lambda_{\mathrm{L}}\leq\lambda_H, A_{\mathrm{L}}\leq A_H} - \lambda_{\mathrm{H}}K \tag{3.132}$$

$$\overset{(3.22)}{\leq} \underbrace{\left[M_{\mathrm{L}}\lambda_L + (M_{\mathrm{H}}-1)\lambda_{\mathrm{H}} - \mu + N\mu\right]}_{=1-\alpha-\lambda_{\mathrm{H}}-\mu,\text{ as } M_{\mathrm{L}}\lambda_L+M_{\mathrm{H}}\lambda_{\mathrm{H}}+N\mu+\alpha=1} \times (A_{\mathrm{H}}-K) + \lambda_{\mathrm{H}}(A_{\mathrm{H}}-K) \tag{3.133}$$

$$= \underbrace{(1-\alpha-\mu)}_{\leq 1,\text{ as } M_{\mathrm{L}}\lambda_L+M_{\mathrm{H}}\lambda_{\mathrm{H}}+N\mu+\alpha=1} \times (A_{\mathrm{H}}-K) \tag{3.134}$$

$$\leq A_{\mathrm{H}} - K. \tag{3.135}$$

Here we use the induction hypothesis and uniformization assumption, similar to Case-1, along with the assumptions that $\lambda_{\mathrm{L}} \leq \lambda_{\mathrm{H}}$ and $A_{\mathrm{L}} \leq A_{\mathrm{H}}$. This concludes the proof of Lemma 5. □

Chapter 4

Optimal Joint Allocation of Control and Communication Channels

Public safety communication systems(PSCSs) are finite source multicast systems. In such systems, the channel allocation is performed on the basis of multicast groups instead of radio units. Thus, in these systems, the call source units are these groups, that we also call the *users*. In a practical system, these users are finite in number and each user or group may consist of many radio units. In this study, we discuss the coupling of uplink control and communication segments (layers) of such finite source systems, due to which the performance of one layer directly impacts that of the other one. However, the conventional theoretical studies model these system segments separately, therefore unable to capture the coupling issues in networks with finite sources' constraint. We first propose a novel model for wireless access systems that incorporates this coupling, by jointly quantifying both the collision loss at the control layer and congestion loss at the communication layer. Under our proposed framework, we further optimize the number of uplink control and communication channels in order to minimize the joint total loss rate given a constraint on the total number of available channels. The optimization results, under all possible traffic parameters needed, are further visualized using our proposed *channel allocation map*. Note that this optimal channel allocation also requires knowledge of the actual call-arrival traffic load. We also demonstrate the capability of our proposed model in estimating the *invisible actual traffic* load, and provide guidelines for developing an algorithm for the *traffic aware* allocation of channels, based on our proposed model. This study has resulted in two research papers, [55, 56].

4.1 Introduction

4.1.1 Motivation

In *multicast* wireless access systems, like trunked mobile radio system [57, 58], radio units are divided into multicast talk groups. In such systems, multicasting generated from within a talk group

is a predominately primary traffic. Also, when a radio talks, the rest of the radios in the group listen, therefore a group busy in communication does not generate a new call request. Therefore, in these systems, the number of groups busy in communications affects the call arrival traffic over the uplink. The future wide-band and nation wide first responder network, the *FirstNet* [59, 60], is one of the modern examples of such systems. In these systems, when a radio needs to talk it sends a call request to the *access point* over an uplink *control channel*, which is dedicated for the *access control* process, e.g., radio A1 of talk group A and radio B4 of talk group B in Figure 4.1. Note

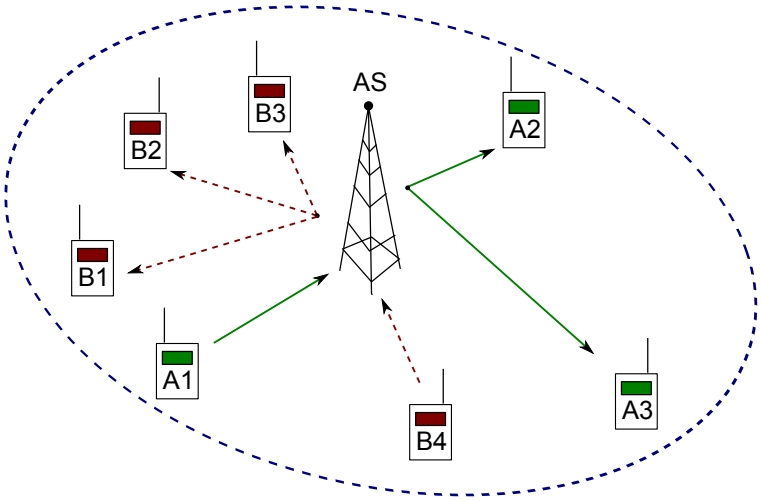


FIGURE 4.1: A multicast system. Two groups A and B, an access station AS.

that the access point is called the access station (AS) in an infrastructure based system [61, 62], and the cluster head in an ad hoc setting [63]. After successfully receiving a call request from a radio, the AS then assigns communication resources to the caller and the rest of the radios of the group, in order to broadcast the call throughout the group, as shown in Figure 4.1. An interesting example of such scenario can be a radio unit of a first responder talk group that shares a video feed of an event's site, with the rest of its talk group members, to share important visual information apart from just voice call [60]. Thus, in such systems, the channel allocation is performed on group basis. Therefore, instead of radios, the *call source units* or the system *'users'* are the groups. Note that these groups or users are also finite in number in practical systems, therefore

we call such systems as *finite source* wireless access systems. However, each user or group may consist of many radios [64].

In a control channel based wireless access system, like the one discussed above, there are two types of frequency channels. One of the types is the *access-control* or simply *control* channel that is dedicated for the *access control* process, whereas the rest of them are of the type of *communication channels*. Therefore we divide the system into two layers, on the basis of both operation and resources, namely, the *access layer* and the *communication layer*. Here the term *layer* does not correspond to the one used in the Open Systems Interconnection (OSI) model, rather it only signifies a particular segment of the system. Note that at the access layer, multiple users may select the same control channel at the same time to send their call requests to the access point, called the access station (AS), which results in contention. Such contention is taken care of by a prescribed multiple access (MAC) protocol [65] through which a subset of contending users get hold of the control channel. As a result, the rest of the failed attempts contributes to the *collision loss*. For those calls that successfully go through the control channel, there are still chances of further losses due to non-availability of channels at the communication layer. Such losses are coined as congestion losses in this study.

Since the bandwidth is a scarce resource, we need to efficiently allocate channels for the access control and communications. Usually in practical systems, only a single frequency channel is dedicated for access control, while the rest of the channels are used for communications. This might not be an optimal allocation strategy, especially in high volume traffic scenarios or in cases where the response time of each accessing unit is short enough to increase the contention rate. In these scenarios, the rate at which multiple users can select the same control channel increases, which increases the collision loss. The collision loss can become so high that even if we have idle communication channels available, the users will not be able to access them due to being frequently blocked at the control layer. A way to alleviate collision is to increase the allowed number of access retrials for every access procedure cycle, which is a standard part of all access

control protocols. But such retrials can only be finite in number and are upper bounded by the delay performance of the access protocol. Thus, for high traffic loads, we might need more than one frequency channels for access control to alleviate collision. However, for a given number of channels available, if we increase the number of control channels to alleviate collisions, we shall be decreasing the number of communication channels at the same time, thereby increasing the congestion loss, and vice-versa. Thus, the channel allocation for access control and communications introduces a trade-off between congestion and collision losses in the system. Hence, the wireless access systems need a mechanism for an optimal channel allocation for the access-control and communications, such that the total system loss is minimized, for a given total number of channels available. This is the problem under consideration in this study. Note that our results, presented in Section 4.6.7, demonstrate that the conventional system with a single control channel performs considerably worse than the system with the same total number of channels but having an optimal allocation for the access control and communications.

What makes such a trade-off issue even more acute is a persistent coupling relationship between two types of losses in the networks with finite sources. This is because the number of control channels and the number of users contending for system access affect the congestion loss by affecting the number of users further demanding for communication channel access, whereas the number of communication channels and the service rate affect the collision loss at the MAC layer by affecting the number of idle and thus potential contending users, given a finite set of users in the system. The latter effect is because in such networks only idle users (groups) can possibly become contending ones. Consequently, it is critical to consider allocating channels to *jointly* cater the needs for both access control and communication services, in terms of minimizing both types of losses while considering the coupling, and a constraint on the total number of channels, for such networks with finite sources. The primary goal of this study is to provide our solutions to such problems. To this end, we propose such a novel model for these systems that jointly

quantifies the collision and congestion losses, and also provides a framework to jointly allocate the control and communication channels to minimize the total system loss.

Another important issue that needs to be resolved is the limitation of using the model for the channel allocation for practical systems. The limitation is that the practical multiple access systems usually only keep records of the calls that successfully get access to the system, but no record is kept for the call requests that are lost due to collision. Hence, for the practical systems, the number of calls that successfully get access and the congestion loss are *visible*, whereas the actual traffic load and the collision loss are *invisible*. Note that the optimal channel allocation requires the knowledge of this invisible actual traffic load. Our proposed model can help overcome this limitation as well, by providing the statistical relationships between the invisible actual traffic load and the *visible* system states, which can be used to estimate the invisible actual traffic load based on the known values of the visible system states. It is this capability of estimating the invisible actual traffic load, provided by our model, which brings the *traffic awareness* characteristic to our proposed channel allocation scheme.

Since collision loss is invisible in practical systems, a possibly misleading performance metric, which we call as the *reported system loss rate*, is often used as the system's performance metric. This reported loss is described in detail in Section 4.5. The reported loss rate is evaluated based on the available system data of the recorded calls and does not count the collision loss. As discussed in Section 4.6.6, collision loss is a significant part of the total loss. In fact, most of the time, it is far more significant than the congestion loss. Therefore, since the reported loss rate does not incorporate the information of the collision loss, it is a considerably underestimated measure of system loss. Thus the reported loss is misleading if we try to make system design decisions on its basis. However, our model resolves this issue by evaluating both the collision and the congestion losses that can be used to evaluate the actual total system loss. This actual total loss should be used instead of the misleading reported loss for making the system design decisions.

4.1.2 Related Works

To the best of our knowledge, the existing works on multicast systems, e.g., [66–72], focus on the *downlink* channels, and on such scenarios wherein the multicast traffic is generated from *outside* the multicast group. Interesting examples of such a scenario are, mobile TV service provided to a group of subscribers, and wireless down-streaming of a shared multimedia content over the Internet. However, in this study, we focus on the *uplink* channels, and on such scenarios wherein the multicast traffic is generated from *within* the multicast group and the number of groups busy in communications affects the call arrival traffic over the uplink.

Recall that in this study, we address the problem of the joint allocation of control and communication channels. To the best of our knowledge, not much literature is available on the joint channel allocation for the access control and communications, while adding the coupled losses over these two types of channels. In most existing works, collision losses in MAC layer and congestion losses in data-link layer have been treated separately. For example, earlier studies like [73, 74], and recent studies, e.g., related to trunked radio systems [75], LTE based systems [76], ad hoc wireless networks [77, 78], underwater acoustic networks [79] and inter-vehicle networks [80, 81]. Note that these studies focus only on the MAC layer, without considering the communication channel allocation and the congestion loss. On the other hand, separate studies on the communication channel allocation are also available, but without considering the control channel allocation, e.g, [42, 43, 82–85]. In these studies, the communication channel allocation is presented as a call admission and/or preemption control, under an assumption that there is no collision in the system. Furthermore, for those works where the finite source systems are studied, e.g., [75], busy sources are still considered to generate Poisson type traffic, thereby without the coupling problem we study here. Therefore, the existing models and approaches cannot completely capture the control-communication coupling that we have addressed in this study and introduced in Section 4.1.1, for the networks with finite sources. The most relevant works to-date is [86] where a simple dynamic algorithm for the allocation of control and communication channels is proposed. But [86] neither

provides any model or analysis for deriving this algorithm, nor does it discuss the optimality of the proposed algorithm. Rather it only demonstrates, via simulation results, that the proposed dynamic algorithm outperforms the fixed channel allocation in case of time varying call traffic. Also, note that in [86], the objective is to improve the channel utilization. However, in this study, our objective is to find the optimal number of control and communication channels that minimizes the total system loss.

4.1.3 Summary of Our Contributions

A summary of our contributions in this study is listed below.

- We propose a novel model for multiple access systems that jointly models the access-control and communication layers. This model helps us quantify both the collision and congestion losses in the system, which are then used to evaluate the *actual total loss rate*. Note that we discuss the model for a simple access control protocol, however, we can easily incorporate more sophisticated protocols by simply updating the access state distribution accordingly.

- We then formulate the optimization problem to jointly find the optimal number of control and communication channels that minimizes the actual total loss rate, for a given number of channels available in total. Due to its complexity, this optimization problem does not have a closed form analytical solution. Therefore, we solve the problem numerically.

- Using numerical results, we demonstrate the existence of the optimal solution. Later, we introduce the concept of a *channel allocation map* to represent the optimal channel allocation for all possible values of the traffic parameters.

- As a further contribution, we use our model to quantify the reported system loss rate and elaborate its misleadingness as compared to the actual loss rate.

- We demonstrate a mechanism, based on our proposed model, which can be used to estimate the invisible actual traffic load, required for deciding the *traffic-aware* optimal channel allocation, and also for estimating the actual loss rate, for a practical system.

- We provide guidelines for developing an algorithm for the traffic aware allocation of control and communication channels, based on our proposed model and its capability to estimate the actual invisible traffic load.

- We demonstrate using numerical results that the optimal channel allocation provides a significant improvement in performance as compared to the conventional strategy of using a single control channel.

- Our results show that even though we spend time resources at the control layer, we may still need more than one frequency channel for access-control, depending on the traffic parameters.

4.1.4 Organization of the Chapter

The rest of the chapter is organized as follows. We present the system model in Section 4.2. The problem formulation is discussed in Section 4.3. A demonstration of estimating the invisible actual traffic load, along with the remarks on the development of the traffic aware channel allocation algorithm, is discussed in Section 4.4. We then discuss the reported loss rate in Section 4.5, and present the numerical and simulation results in Section 4.6. Finally, we conclude in Section 4.7. Also note that for the convenience of the reader, we have presented all the symbols with descriptions in Table 4.1.

4.2 System Model

We consider a single *cell* or a *site* of a multiple access communication system, with a finite number of users (groups), say M, and an *access station* (AS). The access station consists of N frequency channels. Assume that all N channels are of same quality. Out of these N channels, there are $N_{\rm x}$ number of control channels and $N_{\rm c} = N - N_{\rm x}$ number of communication channels. We assume

TABLE 4.1: Table of Symbols

Symbol	Definition
M	Number of users
N	Total number of channels
N_{x}	Number of access control channels
N_{c}	Number of communication channels
N_{xo}	Optimal N_{x}
N_{co}	Optimal N_{c}
I_n	Number of idle users in the nth SP
ω	Communication channel occupancy rate
L_n	Number of contenders in nth AP
X_n	System access state in nth AP
${}_jX_n$	jth access channel state in nth AP
${}_jL_n$	No. of contenders for jth access channel in nth AP
Q_n	Number of calls blocked due to collision in nth SP
Z_n	Number of busy communication channels in nth FP
Y_n	Number of communication channels busy in nth AP
G_n	Number of calls blocked due to congestion in nth SP
Ω_{X}	Sample space of $X_n = \{0, 1, \ldots, N_{\mathrm{x}}\}$
$\Omega_{\mathrm{Y}}(z_1)$	Space of Y_n (given $Z_{n-1} = z_1$) $= \{0, 1, \ldots, z_1\}$
$\overline{\Omega}_{\mathrm{Y}}$	Sample space of $Y_n = \{0, 1, \ldots, N_{\mathrm{c}}\}$
$\Omega_{\mathrm{XY}}(z_1)$	$= \Omega_X \times \Omega_Y(z_1)$
$\overline{\Omega}_{\mathrm{XY}}$	$= \Omega_X \times \overline{\Omega}_Y$
$\Omega_{\mathrm{L}}(z_1)$	Space of L_n (given $Z_{n-1} = z_1$) $= \{0, 1, \ldots, M - z_1\}$
$\overline{\Omega}_{\mathrm{L}}$	Sample space of $L_n = \{0, 1, \ldots, M\}$
Ω_{Z}	Sample space of $Z_n = \{0, 1, \ldots, N_{\mathrm{c}}\}$
Ω_{G}	Sample space of $G_n = \{0, 1, \ldots, N_{\mathrm{x}}\}$
β	Actual total loss rate
β_{Q}	Collision loss rate
β_{G}	Congestion loss rate
$\tilde{\beta}$	Reported loss rate
π_z	$= \lim_{n\to\infty} \Pr(Z_n = z)$
$\underline{\Pi}$	Steady state distribution vector for Z_n
$\mathbf{P}_{\mathrm{z}}$	Transition probability matrix for Z_n
λ	Call arrival rate per idle user
$\hat{\lambda}_n$	Estimated call arrival rate per idle user until nth SP
σ	CRP contending rate per contender
s	Maximum number of access-slots per AP for CRP
$\psi(l_j)$	Pr(access success per access slot for CRP $\Big\vert\ {}_jL_n = l_j$)
$\Psi(l_j)$	Pr(access success for CRP $\Big\vert\ {}_jL_n = l_j$)
Δ_{p}	%age increment in loss rate for $N_{\mathrm{x}} = 1$ w.r.t optimal

that the bandwidths of the control and communication channels are the same. However, in practice, these bandwidth requirements can be different. Note that our proposed model can be easily modified to incorporate this scenario, but such modifications in the model and the corresponding analysis are left as a part of future work. The system employs a *synchronized random access protocol*, e.g. slotted-Aloha [87, 88]. Under this protocol, in our model, the access procedure is synchronized and discrete in time. Also, a user only attempts for the system access at the beginning of an *access time slot*. During the same time slot the user gets a reply from the AS regarding the success or failure of the access attempt. Note that we consider a loss system, i.e., there is no queueing buffer for the calls. This assumption helps analyze and highlight the coupling of the control and communication layers, and its impact on the allocation of channels, which is the main objective of this study, without making the model too complicated.

We model the system as a discrete time Markov chain (DTMC), with the time horizon divided into discrete time segments, called the system periods (SPs). We assume that the time scale of SP is such that each idle user (group) can have at the most one call arrival during an SP. Each system period is divided into two time segments as shown in Figure 4.2. The time segment, at

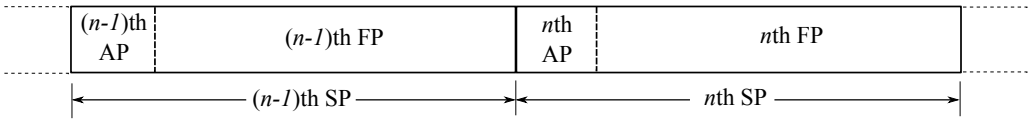


FIGURE 4.2: Time-horizon discretization.

the beginning of an SP, during which all the access process takes place, is called the *access period* (AP). During an AP, all the contending users try to access the system and the AS replies to their requests. At the end of an AP the communication channel allocation occurs that marks the beginning of the *access free period* (FP) in the system period. Therefore, an AP can also be considered as the pre communication channel allocation segment of an SP, whereas an FP can be considered as the post communication channel allocation segment. Note that a similar time segmentation is used in [89] as well.

We present a graphical representation of our system model in Figure 4.3. Consider an nth SP and suppose there were I_{n-1} number of *idle users* during the $(n-1)$th SP. These are the users

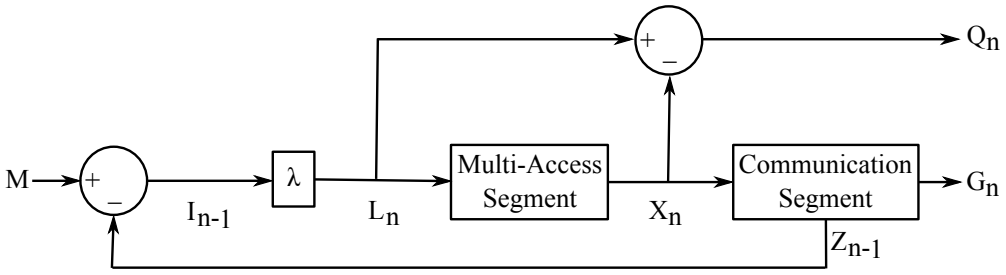


FIGURE 4.3: System model.

which were not busy in communications, and thus not occupying the communication channels, during the $(n-1)$th FP. Out of these I_{n-1} idle users, some or all users will contend for system access in the next slot, i.e., the nth SP. Let L_n be the number of contenders during the nth SP, that decide to contend out of I_{n-1} idle users. Out of these L_n contenders, some users successfully get access to the system while the rest of the contenders are blocked due to collision. The details of the access procedure considered for this study are provided in the next paragraph. In Section 4.3.4 we also describe how to incorporate any other access control protocol in the model.

Access control procedure: Assume that a call arrives for an idle user during an $(n-1)$th SP. The user then contends for the system access in the nth AP, by sending a call request to the AS over a randomly selected control channel. A call request over a control channel successfully gets access to the system to enter the communication channel allocation process, if and only if, only one contender selects that channel. Otherwise, if two or more contenders select the same control channel, in the same AP then all such call requests are blocked due to collision, and the collided users immediately go into a *collision resolution procedure* (CRP) mode, explained in the next paragraph. We assume that a contending user does not change the selected control channel for the CRP mode. If an access success does not occur, even during the CRP, the call is blocked and is counted as a collision loss for the user. The user then immediately goes to the idle state.

Collision resolution procedure (CRP): We consider a simple CRP in which each contender either contends for the system access with probability $\sigma \in (0, 1)$, at the beginning of each access slot during an AP, or it waits during that access slot with probability $1 - \sigma$. We call σ as the *CRP contending rate per contender*, and it is same for all the users. Also, we assume that an AP can have at the most $s \in \mathbb{N}^+$ access slots for CRP. Therefore, s is the maximum number of access retrials per user allowed during an AP. Thus, the CRP ends either when an access success occurs, or after s access slots in an AP. Note that an access success occurs for an access control channel when exactly one user selects that channel during a given access slot. This CRP is a special case of the standard *collision avoidance* technique, where the *random back off* method is employed [90]. In our CRP, the back off time, in terms of the number of access slots, is geometrically distributed with parameter $1 - \sigma$.

Let X_n be the number of users or calls successfully getting access to the system out of L_n contenders and $Q_n = L_n - X_n$ be the number of users or call requests blocked due to collision, during the nth AP. We also call X_n as the *system access state*. Out of X_n users successfully getting access to the system, some of the users are allocated communication channels at the beginning of the nth FP depending on the number of communication channels available. While the rest of the users are blocked due to congestion, i.e., unavailability of communication channels. Let Z_n be the number of busy communication channels during the nth FP. Note that the number of busy communication channels and that of users or calls busy in communications are the same. Then Z_{n-1} will be the number of busy communication channels during the $(n-1)$th FP. Out of these Z_{n-1} communication channels, some number of channels, say Y_n, still remain busy during the nth AP and FP. Thus, Y_n is also the number of communication channels monitored busy during the nth AP, before the communication channel allocation for the nth SP. Let ω be the probability that a communication channel that is busy in a given FP remains busy for the next FP as well. We call ω as the *communication channel occupancy rate* per busy user. We assume that ω is same for

all the users and communication channels. Given Z_{n-1}, Y_n is binomially distributed as follows:

$$\Pr\big(Y_n = y \big| Z_{n-1} = z_1\big) = \begin{cases} \binom{z_1}{y}\omega^y(1-\omega)^{z_1-y} & \text{if } y \in \Omega_{\mathrm{Y}}(z_1), \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

$\forall y \in \overline{\Omega}_{\mathrm{Y}}$ and $\forall z_1 \in \Omega_{\mathrm{Z}}$. Here $\Omega_{\mathrm{Y}}(z_1) = \{0, 1, \ldots, z_1\}$ and $\overline{\Omega}_{\mathrm{Y}} = \Omega_{\mathrm{Z}} = \{0, 1, \ldots, N_{\mathrm{c}}\}$. Note that here we assume a memoryless and stationary discrete time call-service process. It is similar to the commonly used exponential service time process, which is also memoryless but continuous in time, as in [41–43]. In practical systems, the quality of the downlink channels also affect ω. However, we leave the detailed modeling of the downlink for future work, and in this study, only consider the coupling between the control and communication layers over the uplink. For this study, we consider ω as a known traffic parameter that can be evaluated from practical system data of call records.

Communication channel allocation: The number of idle communication channels during the nth AP is $N_{\mathrm{c}} - Y_n$, before the allocation of communication channels to the new calls. If $X_n \leq N_{\mathrm{c}} - Y_n$, i.e., the number of calls successfully getting access to the system is less than or equal to the number of idle communication channels, then there will be no congestion. In this case, all new X_n calls are allocated communication channels without any call blocking, making $Z_n = X_n + Y_n$. Otherwise, $X_n > N_{\mathrm{c}} - Y_n$ and the system randomly selects $N_{\mathrm{c}} - Y_n$ calls out of X_n new calls, making all the channels busy during the nth FP and thus $Z_n = N_{\mathrm{c}}$. We call this a *random call (or user) selection*. The remaining $X_n - (N_{\mathrm{c}} - Y_n)$ calls are blocked due to congestion. Thus, we can write the number of busy communication channels Z_n in terms of the X_n and Y_n as follows:

$$Z_n = \begin{cases} X_n + Y_n & \text{if } X_n + Y_n \leq N_{\mathrm{c}}, \\ N_{\mathrm{c}} & \text{otherwise.} \end{cases} \tag{4.2}$$

Let G_n be the calls lost due to congestion during an nth SP. We can also write G_n in terms of the X_n and Y_n as follows:

$$G_n = \begin{cases} 0 & \text{if } X_n + Y_n \leq N_{\mathrm{c}}, \\ X_n + Y_n - N_{\mathrm{c}} & \text{otherwise.} \end{cases} \tag{4.3}$$

Note that in finite source wireless systems, e.g., trunked radio systems, a talk group busy in communications does not generate a new call request. Thus, only an idle user, not busy in communications, can generate a new call request and contend for system access. The number of idle users during $(n-1)$th SP is $I_{n-1} = M - Z_{n-1}$, where Z_{n-1} is the number of channels, or equivalently users, busy in communications during the $(n-1)$th FP. Out of these I_{n-1} idle users, L_n users will contend for system access in the next slot, i.e., the nth SP. Let λ be the probability that a call arrives for an idle user during an SP. We call λ as the *call arrival rate per idle user*, and assume that it is same for all the users in the system. Now L_n is binomially distributed and depends on the number of idle users $M - Z_{n-1}$ in the $(n-1)$th SP, i.e:

$$\Pr\big(L_n = l \big| Z_{n-1} = z_1\big) = \begin{cases} \binom{M-z_1}{l} \lambda^l (1-\lambda)^{M-z_1-l} & \text{if } l \in \Omega_{\mathrm{L}}(z_1), \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

$\forall l \in \overline{\Omega}_{\mathrm{L}}, \forall z_1 \in \Omega_{\mathrm{Z}}$. Here, $\Omega_{\mathrm{L}}(z_1) = \{0, 1, \ldots, M - z_1\}$ and $\overline{\Omega}_{\mathrm{L}} = \{0, 1, \ldots, M\}$. Note that for a sufficiently large M and small λ, the binomial traffic model (4.4) approaches the Poisson model, which is commonly used for analysis of communication systems, e.g. in [41–43]. Also note that $\Pr\big(L_n = l \big| Z_{n-1} = z_1\big)$ is a function of L_n, Z_{n-1}, M and λ, and can be easily updated to incorporate any other call traffic model of interest.

As explained in the previous paragraph, the call traffic model (4.4) models the arrival of calls such that the new calls are generated only by the idle users that are not busy in communications. This also agrees with the operation of practical finite source systems, with both unicast and multicast traffics, because, in such systems, a user busy in communications does not generate a new call request. Due to the same phenomenon, the collision loss at the control layer is affected by the

communication layer performance, as discussed in Section 4.1. Note that we cannot capture this coupling between control and communication layers without such a traffic model. However, this dependence of call arrival traffic on the number of idle users is not considered in conventional studies related to unicast and multicast finite source systems, e.g, in [69, 72, 91].

In our study, we assume that the time horizon begins at $n = 0$, with the initial values of all the processes being 0, i.e., $L_0 = 0$, $X_0 = 0$, $Y_0 = 0$, $Q_0 = 0$ and $G_0 = 0$, with probability 1. Since $X_0 = 0$ and $Y_0 = 0$, therefore $Z_0 = 0$ according to (4.2). Moreover, all the expressions in this study are for $n \in \mathbb{N}^+$, where $\mathbb{N}^+$ is a set of all strictly positive natural numbers, i.e., natural numbers without 0.

As described so far, and also illustrated in Figure 4.3; L_n, X_n, Y_n, Q_n and G_n depend on Z_{n-1}. Also, given Z_{n-1}; L_n, X_n, Y_n, and therefore Q_n and G_n are stationary processes. Furthermore, given Z_{n-1}; $\{L_n\}_{n=1}^{\infty}$, $\{X_n\}_{n=1}^{\infty}$, $\{Y_n\}_{n=1}^{\infty}$, and therefore $\{Q_n\}_{n=1}^{\infty}$ and $\{G_n\}_{n=1}^{\infty}$ are sequences of *conditionally independent and identically distributed* random variables. Moreover, (4.2) shows that Z_n depends on X_n and Y_n both of which, in turn, depend on Z_{n-1}. Thus, Z_n depends on Z_{n-1}, $\forall n \in \mathbb{N}^+$. Also, Z_n depends on its previous history but only through Z_{n-1}, i.e., Z_n does not depend on the rest of the past values given Z_{n-1}. Therefore, the process Z_n forms a first order discrete time Markov chain in our model.

Remarks on Visibility of Parameters: In practice, multiple access systems only keep records of the calls that successfully get access to the system. But no record is kept for the call requests that are lost due to collision. Therefore, we classify our model parameters, also illustrated in Figure 4.3, into two types. The first type of parameters are those which are recorded by or known to the practical systems, namely, ω, X_n, Z_n, G_n, M and N. We call them the *visible* parameters. The second type of parameters are those which are not recorded by or unknown to the practical systems, namely, λ, L_n and Q_n. We call them the *invisible* parameters. The key invisible parameter is λ and we also call it the *invisible actual traffic load.* In Section 4.4, we demonstrate how we can use our model to estimate λ, using the known values of the visible parameters, under the model

proposed here. Note that this capability of estimating λ provided by our model brings the *traffic awareness* characteristic to our proposed channel allocation scheme, as described in Section 4.4. Also, in Section 4.5.1, we remark how we can acquire the required knowledge of L_n and Q_n based on the estimated value of λ.

4.3 Problem Formulation

4.3.1 Performance Metric

The performance metric in our study is the *total loss rate* of the system. We define it as the *fraction of calls blocked* during an SP in the long run, i.e., infinite time horizon or $n \to \infty$. It is represented as β and given by:

$$\beta = \lim_{k\to\infty} \frac{\sum_{n=1}^{k} Q_n + G_n}{\sum_{n=1}^{k} L_n} = \beta_{\mathrm{Q}} + \beta_{\mathrm{G}}. \tag{4.5}$$

Here β_{Q} and β_{G} are the collision and congestion loss rates of the system respectively, explained in the following paragraphs. Note that for a given channel allocation and the number of contenders in the system, the collision and congestion are independent loss events.

Collision Loss Rate: We define the collision loss rate β_{Q} as the *fraction of calls blocked due to collision*, during an SP in the long run. Mathematically, irrespective of any assumed access control and communication channel allocation protocol, for $\sum_{\forall z_1 \in \Omega_{\mathrm{Z}}} \mathrm{E}\big[L_n \big| Z_{n-1} = z_1\big].\pi_{z_1} \neq 0$:

$$\beta_{\mathrm{Q}} = \lim_{k\to\infty} \frac{\sum_{n=1}^{k} Q_n}{\sum_{n=1}^{k} L_n} \tag{4.6}$$

$$= \frac{\sum_{\forall z_1 \in \Omega_{\mathrm{Z}}} \mathrm{E}\big[Q_n \big| Z_{n-1} = z_1\big].\pi_{z_1}}{\sum_{\forall z_1 \in \Omega_{\mathrm{Z}}} \mathrm{E}\big[L_n \big| Z_{n-1} = z_1\big].\pi_{z_1}}. \tag{4.7}$$

Here $\Omega_{\mathrm{Z}} = \{0, 1, \ldots, N_{\mathrm{c}}\}$ is the sample space of Z_n, E[.] is the expectation operator and π_{z_1} is the steady state probability of the Markov chain Z_n for state $Z_n = z_1$, i.e., $\lim_{n\to\infty} \Pr(Z_n = z_1)$. Note that we get (4.7) from (4.6) by using the *Law of Large Numbers*, and the fact that given Z_{n-1}, both $\big\{L_n\big\}_{n=1}^{\infty}$ and $\big\{Q_n\big\}_{n=1}^{\infty}$ are sequences of *conditionally independent and identically distributed* random variables. The steady state probability π_{z_1} is presented in Section 4.3.2. The numerator in (4.7) is the average number of calls blocked due to collision during an SP, over

the long run. While the denomenator is the average number of contenders or arrived calls in an SP, over the long run. Since, given Z_{n-1}, L_n is binomially distributed as shown in (4.4), therefore $\mathrm{E}\big[L_n\big|Z_{n-1}=z_1\big]=(M-z_1)\lambda$. Also, we know that $Q_n=L_n-X_n$, therefore, $\mathrm{E}\big[Q_n\big|Z_{n-1}=z_1\big]$ $=\mathrm{E}\big[L_n\big|Z_{n-1}=z_1\big]-\mathrm{E}\big[X_n\big|Z_{n-1}=z_1\big]$. If we represent the sample space of X_n as the set Ω_{X}, $\Omega_{\mathrm{X}}=\{0,1,\ldots,N_x\}$, then we have $\mathrm{E}\big[X_n\big|Z_{n-1}=z_1\big]=\sum_{\forall x\in\Omega_{\mathrm{X}}}x\mathrm{Pr}\big(X_n=x\big|Z_{n-1}=z_1\big)$. We call $\mathrm{Pr}\big(X_n=x\big|Z_{n-1}=z_1\big)$ as the *conditional distribution of the system access state*, and is presented in Section 4.3.4.

Congestion Loss Rate: We define the congestion loss rate β_{G} as the *fraction of calls blocked due to congestion*, during an SP in the long run. Mathematically, similar to Equation (4.7), for $\sum_{\forall z_1\in\Omega_{\mathrm{Z}}}\mathrm{E}\big[L_n\big|Z_{n-1}=z_1\big].\pi_{z_1}\neq 0$:

$$\beta_{\mathrm{G}}=\lim_{k\to\infty}\frac{\sum_{n=1}^{k}G_n}{\sum_{n=1}^{k}L_n} \tag{4.8}$$

$$=\frac{\sum_{\forall z_1\in\Omega_{\mathrm{Z}}}\mathrm{E}\big[G_n\big|Z_{n-1}=z_1\big].\pi_{z_1}}{\sum_{\forall z_1\in\Omega_{\mathrm{Z}}}\mathrm{E}\big[L_n\big|Z_{n-1}=z_1\big].\pi_{z_1}}. \tag{4.9}$$

The numerator in (4.9) is the average number of calls blocked due to congestion during an SP, over the long run. If we represent the sample space of G_n as $\Omega_{\mathrm{G}}=\{0,1,\ldots,N_x\}$, then we have $\mathrm{E}\big[G_n\big|Z_{n-1}=z_1\big]=\sum_{\forall g\in\Omega_{\mathrm{G}}}g\,.\,\mathrm{Pr}\big(G_n=g\big|Z_{n-1}=z_1\big)$. We call $\mathrm{Pr}\big(G_n=g\big|Z_{n-1}=z_1\big)$ as the *conditional distribution of the congestion loss*, and is presented in Section 4.3.3.

4.3.2 Steady State Distribution of Z_n

As mentioned in Section 4.2, Z_n forms a discrete time Markov chain (DTMC). Also, this chain is irreducible and ergodic. Therefore a steady state distribution exists for Z_n. Let π_z be the steady state probability of the Markov chain Z_n for state $Z_n=z$, i.e., $\pi_z=\lim_{n\to\infty}\mathrm{Pr}(Z_n=z)$. Let $\underline{\Pi}$ be a column vector such that its zth element is π_z. Then the steady state distribution is the solution of a system of linear equations, i.e., $\underline{\Pi}=\mathbf{P}_{\mathrm{z}}^{\mathrm{t}}\,\underline{\Pi}$ and $\sum_{\forall z\in\Omega_{\mathrm{Z}}}\pi_z=1$. Here $\mathbf{P}_{\mathrm{z}}^{\mathrm{t}}$ is the transpose of matrix $\mathbf{P}_{\mathrm{z}}$ which is the transition probability matrix for DTMC Z_n. Thus, the element of matrix $\mathbf{P}_{\mathrm{z}}$ at z_1th row and z_2th column is the state transition probability, i.e., $\mathbf{P}_{\mathrm{z}}(z_1,z_2)=\mathrm{Pr}\big(Z_n=z_2\big|Z_{n-1}=z_1\big)$, $\forall z_1,z_2\in\Omega_{\mathrm{Z}}$. This transition probability can be evaluated

by marginalizing the conditional joint distribution $\Pr\big(Z_n = z_2, X_n = x, Y_n = y \big| Z_{n-1} = z_1\big)$, as shown in the following expression.

$$\Pr\big(Z_n = z_2 \big| Z_{n-1} = z_1\big) = \sum_{\forall (x,y) \in \overline{\Omega}_{\mathrm{XY}}} \Pr\big(Z_n = z_2, X_n = x, Y_n = y \big| Z_{n-1} = z_1\big), \tag{4.10}$$

$\forall z_1, z_2 \in \Omega_{\mathrm{Z}}$. Also, $\overline{\Omega}_{\mathrm{XY}} = \Omega_{\mathrm{X}} \times \overline{\Omega}_{\mathrm{Y}}$. The summand in (4.10) can be broken down using the *chain rule of probability* and the facts that given X_n and Y_n, Z_n is independent of Z_{n-1}, and given Z_{n-1}, Y_n is independent of X_n:

$$\begin{aligned}
&\Pr\big(Z_n = z_2, X_n = x, Y_n = y \big| Z_{n-1} = z_1\big) \\
&= \Pr\big(Z_n = z_2 \big| Y_n = y, X_n = x\big) \times \Pr\big(Y_n = y \big| Z_{n-1} = z_1\big) \times \Pr\big(X_n = x \big| Z_{n-1} = z_1\big),
\end{aligned} \tag{4.11}$$

$\forall x \in \Omega_{\mathrm{X}}$, $\forall y \in \overline{\Omega}_{\mathrm{Y}}$ and $\forall z_1, z_2 \in \Omega_{\mathrm{Z}}$.

Consider the first term in (4.11). It is given by the following equation, based on our knowledge of (4.2), wherein we express Z_n in terms of X_n and Y_n.

$$\begin{aligned}
&\Pr\big(Z_n = z_2 \big| Y_n = y, X_n = x\big) \\
&= \begin{cases} 1 & \text{if } (x + y \leq N_{\mathrm{c}} \text{ and } z_2 = x + y) \text{ or } (x + y > N_{\mathrm{c}} \text{ and } z_2 = N_{\mathrm{c}}), \\ 0 & \text{otherwise,} \end{cases}
\end{aligned} \tag{4.12}$$

$\forall x \in \Omega_{\mathrm{X}}$, $\forall y \in \overline{\Omega}_{\mathrm{Y}}$ and $\forall z_1, z_2 \in \Omega_{\mathrm{Z}}$. The second term in (4.11), $\Pr\big(Y_n = y \big| Z_{n-1} = z_1\big)$, is given by (4.1), whereas the last term, $\Pr\big(X_n = x \big| Z_{n-1} = z_1\big)$, is presented in Section 4.3.4. Thus, the summand in (4.10) can be evaluated using (4.11) which can help us evaluating the transition probabilities using (4.10). These transition probabilities are then used to form the transition probability matrix $\mathbf{P}_{\mathrm{z}}$ which in turn is used to evaluate the steady state distribution, π_z $\forall z \in \Omega_{\mathrm{Z}}$. Recall that these steady state distributions are required for evaluating the collision and congestion loss rates, using (4.7) and (4.9) respectively.

4.3.3 Conditional Distribution of the Congestion Loss G_n

Now we evaluate the conditional distribution of the congestion loss G_n, $\Pr\big(G_n = g \big| Z_{n-1} = z_1\big)$. For this, we use the same approach as that used in Section 4.3.2. To this end, we marginalize

the conditional joint distribution $\Pr\big(G_n = g, X_n = x, Y_n = y \big| Z_{n-1} = z_1\big)$ as demonstrated in the following mathematical expression.

$$\Pr\big(G_n = g \big| Z_{n-1} = z_1\big) = \sum_{\forall (x,y) \in \overline{\Omega}_{\mathrm{XY}}} \Pr\big(G_n = g, X_n = x, Y_n = y \big| Z_{n-1} = z_1\big), \tag{4.13}$$

$\forall g \in \Omega_{\mathrm{G}}, \forall z_1 \in \Omega_{\mathrm{Z}}$. The summand in (4.13) can be broken down similar to (4.11), and thus given by the following expression.

$$\begin{aligned} &\Pr\big(G_n = g, X_n = x, Y_n = y \big| Z_{n-1} = z_1\big) \\ &= \Pr\big(G_n = g \big| Y_n = y, X_n = x\big) \times \Pr\big(Y_n = y \big| Z_{n-1} = z_1\big) \times \Pr\big(X_n = x \big| Z_{n-1} = z_1\big), \end{aligned} \tag{4.14}$$

$\forall x \in \Omega_{\mathrm{X}}, \forall y \in \overline{\Omega}_{\mathrm{Y}}, \forall g \in \Omega_{\mathrm{G}}$ and $\forall z_1 \in \Omega_{\mathrm{Z}}$.

Consider the first term in (4.14). It is given by the following equation, based on our knowledge of (4.3), wherein we express G_n in terms of X_n and Y_n.

$$\begin{aligned} &\Pr\big(G_n = g \big| Y_n = y, X_n = x\big) \\ &= \begin{cases} 1 & \text{if } (x + y \leq N_{\mathrm{c}} \text{ and } g = 0) \text{ or } (x + y > N_{\mathrm{c}} \text{ and } g = x + y - N_{\mathrm{c}}), \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{4.15}$$

$\forall x \in \Omega_{\mathrm{X}}, \forall y \in \overline{\Omega}_{\mathrm{Y}}, \forall z_1 \in \Omega_{\mathrm{Z}}$ and $\forall g \in \Omega_{\mathrm{G}}$. Similar to Section 4.3.2, the second term in (4.14) is given by (4.1), whereas the last term, $\Pr\big(X_n = x \big| Z_{n-1} = z_1\big)$, is presented in Section 4.3.4. Thus, the summand in (4.13) can be evaluated using (4.14) which can help us evaluate the required probability, $\Pr\big(G_n = g \big| Z_{n-1} = z_1\big)$, $\forall g \in \Omega_{\mathrm{G}}$ and $\forall z_1 \in \Omega_{\mathrm{Z}}$, using (4.13). Recall that this conditional distribution of the congestion loss G_n is used to evaluate the expected congestion loss per SP conditioned on Z_{n-1}, i.e., $\mathrm{E}\big[G_n \big| Z_{n-1} = z_1\big] = \sum_{\forall g \in \Omega_{\mathrm{G}}} g \,.\, \Pr\big(G_n = g \big| Z_{n-1} = z_1\big)$, as mentioned in Section 4.3.1. This expected congestion loss is required for evaluating the congestion loss rate, as shown in (4.9).

4.3.4 Conditional Distribution of Access State X_n

Now, in the context of the access control procedure described in Section 4.2, we evaluate the conditional distribution of X_n, given Z_{n-1}. This distribution can be evaluated by marginalizing the

conditional joint distribution $\Pr\big(X_n = x, L_n = l \big| Z_{n-1} = z_1\big)$, as demonstrated by the following mathematical expression.

$$\Pr\big(X_n = x \big| Z_{n-1} = z_1\big) = \sum_{\forall l \in \overline{\Omega}_L} \Pr\big(X_n = x, L_n = l \big| Z_{n-1} = z_1\big), \tag{4.16}$$

$\forall x \in \Omega_{\mathrm{X}}$ and $\forall z_1 \in \Omega_{\mathrm{Z}}$. The summand in (4.16) can be broken down using the chain rule of probability and the fact that given L_n, X_n is independent of Z_{n-1}, also shown by Figure 4.3. Thus, we can write,

$$\Pr\big(X_n = x, L_n = l \big| Z_{n-1} = z_1\big) = \Pr\big(X_n = x \big| L_n = l\big) \times \Pr\big(L_n = l \big| Z_{n-1} = z_1\big), \tag{4.17}$$

$\forall x \in \Omega_{\mathrm{X}}$, $\forall l \in \overline{\Omega}_{\mathrm{L}}$ and $\forall z_1 \in \Omega_{\mathrm{Z}}$. The second product term in (4.17), $\Pr\big(L_n = l \big| Z_{n-1} = z_1\big)$, is given by (4.4), and in Section 4.3.5 we describe how to evaluate the first term, $\Pr\big(X_n = x \big| L_n = l\big)$. Note that in Section 4.3.5 we present the details according to the assumed access protocol. Also, it is only $\Pr\big(X_n = x \big| L_n = l\big)$ which needs to be updated if the protocol changes, the rest of the framework remains unchanged. This concludes the evaluation of $\Pr\big(X_n = x \big| Z_n = z_1\big)$. Recall that we need this conditional distribution of the system access state for the analysis discussed so far. Also note that the evaluation of this distribution requires values of the model parameters, namely, M, N_{x}, N_{c}, s, σ, λ and ω.

4.3.5 Evaluation of $\Pr\big(X_n = x \big| L_n = l\big)$

We consider three cases for all possible values of L_n, namely $L_n = 0$, $L_n = 1$ and $L_n \geq 2$. When $L_n = 0$, there is no contender in the nth AP. Thus, in that AP, no call gets a successful access to the system and $X_n = 0$ with probability 1, so $\Pr\big(X_n = 0 \big| L_n = 0\big) = 1$. Now consider the case when $L_n = 1$. In this case, there is only one contender in the nth AP, and thus no collision. Therefore, in that AP, this single contender will surely get a successful access to the system, due to no collision and $X_n = 1$ with probability 1, so $\Pr\big(X_n = 1 \big| L_n = 1\big) = 1$. For the case when $L_n \geq 2$, we evaluate the access state probability using a general framework based on the assumed *access control procedure*. To this end, firstly, we consider a binary state variable, ${}_jX_n \in \{0, 1\}$,

called the *jth access channel state* in the nth AP. This access channel state is defined by the following mathematical expression.

$$ {}_jX_n = \begin{cases} 1 & \text{if } j\text{th access channel has an access success in } n\text{th AP}, \\ 0 & \text{otherwise}. \end{cases} \tag{4.18} $$

Thus, $X_n = \sum_{j=1}^{N_{\mathrm{x}}} {}_jX_n$. Let $\left({}_jX_n\right)_{j=1}^{N_{\mathrm{x}}} \in \{0,1\}^{N_{\mathrm{x}}}$ represent an N_{x}-dimensional random-vector, and $\left(x_j\right)_{j=1}^{N_{\mathrm{x}}} = x_1^{N_{\mathrm{x}}} \in \{0,1\}^{N_{\mathrm{x}}}$ represent an N_{x}-dimensional deterministic-vector. For each $x \in \Omega_{\mathrm{X}}$, we define a set $\mathcal{S}_{\mathrm{X}}(x) \subset \{0,1\}^{N_{\mathrm{x}}}$ as follows:

$$ \mathcal{S}_{\mathrm{X}}(x) = \left\{ x_1^{N_{\mathrm{x}}} \in \{0,1\}^{N_{\mathrm{x}}} : \sum_{j=1}^{N_{\mathrm{x}}} x_j = x \right\}. \tag{4.19} $$

Now, the access state distribution, given L_n, can be expressed as:

$$ \begin{aligned} \Pr\left(X_n = x \middle| L_n = l\right) &= \Pr\left(\bigcup_{x_1^{N_{\mathrm{x}}} \in \mathcal{S}_{\mathrm{X}}(x)} \{ \left({}_jX_n\right)_{j=1}^{N_{\mathrm{x}}} = x_1^{N_{\mathrm{x}}} \} \;\middle|\; L_n = l \right) \\ &= \sum_{\forall x_1^{N_{\mathrm{x}}} \in \mathcal{S}_{\mathrm{X}}(x)} \Pr\left(\left({}_jX_n\right)_{j=1}^{N_{\mathrm{x}}} = x_1^{N_{\mathrm{x}}} \middle| L_n = l \right), \end{aligned} \tag{4.20} $$

$\forall x \in \Omega_{\mathrm{X}}$. Note that this expression is a general expression, valid for all $l \in \overline{\Omega}_{\mathrm{L}} = \{0, 1, \ldots, M\}$. But, since for $l = 0$ and $l = 1$ we already know the access state distribution, therefore we only use this general expression for $l \geq 2$. Hence, we can summarize the distribution of the access state, given L_n, as follows:

$$ \begin{aligned} &\Pr\left(X_n = x \middle| L_n = l\right) \\ &= \begin{cases} 1 & \text{if } (x = 0 \text{ and } l = 0) \text{ or } (x = 1 \text{ and } l = 1), \\ \sum_{\forall x_1^{N_{\mathrm{x}}} \in \mathcal{S}_{\mathrm{X}}(x)} \Pr\left(\left({}_jX_n\right)_{j=1}^{N_{\mathrm{x}}} = x_1^{N_{\mathrm{x}}} \middle| L_n = l \right) & \text{if } l \geq 2, \\ 0 & \text{otherwise}, \end{cases} \end{aligned} \tag{4.21} $$

$\forall x \in \Omega_{\mathrm{X}}$ and $\forall l \in \overline{\Omega}_{\mathrm{L}}$. In the following discussion we shall discuss how to evaluate the summand in (4.21) for the case of $L_n \geq 2$.

Consider an nth access period AP. Assume that there are two or more contenders in this AP contending for system access, i.e., $L_n \geq 2$. According to the access procedure, each contender randomly selects an access control channel, out of $N_{\rm x}$ available access channels. Let ${}_jL_n$ be the number of contenders selecting jth access channel, during the nth AP. Thus, $L_n = \sum_{j=1}^{N_{\rm x}} {}_jL_n$. Let $\big({}_jL_n\big)_{j=1}^{N_{\rm x}} \in (\overline{\Omega}_{\rm L})^{N_{\rm x}}$ represent an $N_{\rm x}$-dimensional random-vector, and $\big(l_j\big)_{j=1}^{N_{\rm x}} = l_1^{N_{\rm x}} \in (\overline{\Omega}_{\rm L})^{N_{\rm x}}$ represent an $N_{\rm x}$-dimensional deterministic-vector. For each $l \in \overline{\Omega}_{\rm L}$, we define a set $\mathcal{S}_{\rm L}(l) \subset (\overline{\Omega}_{\rm L})^{N_{\rm x}}$ as follows:

$$\mathcal{S}_{\rm L}(l) = \Big\{ l_1^{N_{\rm x}} \in (\overline{\Omega}_{\rm L})^{N_{\rm x}} : \sum_{j=1}^{N_{\rm x}} l_j = l \Big\}. \tag{4.22}$$

Now, the summand in (4.21), for the case $l \geq 2$, can be evaluated by marginalizing a joint distribution as follows:

$$\begin{aligned} &\Pr\Big(\big({}_jX_n\big)_{j=1}^{N_{\rm x}} = x_1^{N_{\rm x}} \Big| L_n = l \Big) \\ &= \sum_{\forall l_1^{N_{\rm x}} \in \mathcal{S}_{\rm L}(l)} \Pr\Big(\big({}_jX_n\big)_{j=1}^{N_{\rm x}} = x_1^{N_{\rm x}}, \big({}_jL_n\big)_{j=1}^{N_{\rm x}} = l_1^{N_{\rm x}} \Big| L_n = l \Big), \end{aligned} \tag{4.23}$$

$\forall x_1^{N_{\rm x}} \in \mathcal{S}_{\rm X}(x)$ and $l \geq 2$. Now the summand in (4.23) can be broken down using the chain rule of probability as follows:

$$\begin{aligned} &\Pr\Big(\big({}_jX_n\big)_{j=1}^{N_{\rm x}} = x_1^{N_{\rm x}}, \big({}_jL_n\big)_{j=1}^{N_{\rm x}} = l_1^{N_{\rm x}} \Big| L_n = l \Big) \\ &= \Pr\Big(\big({}_jX_n\big)_{j=1}^{N_{\rm x}} = x_1^{N_{\rm x}} \Big| \big({}_jL_n\big)_{j=1}^{N_{\rm x}} = l_1^{N_{\rm x}}, L_n = l \Big) \times \Pr\Big(\big({}_jL_n\big)_{j=1}^{N_{\rm x}} = l_1^{N_{\rm x}} \Big| L_n = l \Big) \qquad (4.24) \\ &= \Pr\Big(\big({}_jX_n\big)_{j=1}^{N_{\rm x}} = x_1^{N_{\rm x}} \Big| \big({}_jL_n\big)_{j=1}^{N_{\rm x}} = l_1^{N_{\rm x}} \Big) \times \Pr\Big(\big({}_jL_n\big)_{j=1}^{N_{\rm x}} = l_1^{N_{\rm x}} \Big| L_n = l \Big), \qquad (4.25) \end{aligned}$$

$\forall x_1^{N_{\rm x}} \in \mathcal{S}_{\rm X}(x)$, $\forall l_1^{N_{\rm x}} \in \mathcal{S}_{\rm L}(l)$ and $l \geq 2$. Note that the first product term in (4.24) is equal to that in (4.25) because for $l_1^{N_{\rm x}} \in \mathcal{S}_{\rm L}(l)$, the events $\big({}_jL_n\big)_{j=1}^{N_{\rm x}} = l_1^{N_{\rm x}}$ and $L_n = l$ are equivalent, as shown by the definition of set $\mathcal{S}_{\rm L}(l)$ in (4.22). Recall that according to the access procedure, each contender randomly selects an access control channel out of $N_{\rm x}$ available channels. Also, all $N_{\rm x}$ control channels are equally likely to be selected by each user. Thus, given the number of contenders

in the nth SP, L_n, $\big({}_jL_n\big)_{j=1}^{N_{\rm x}}$ follows a multinomial distribution, i.e.,

$$\Pr\Big(\big({}_jL_n\big)_{j=1}^{N_{\rm x}} = l_1^{N_{\rm x}}\Big|L_n = l\Big) = \frac{l!}{\prod_{j=1}^{N_{\rm x}} l_j!}(\frac{1}{N_{\rm x}})^l, \tag{4.26}$$

$\forall l_1^{N_{\rm x}} \in \mathcal{S}_{\rm L}(l)$, $l \geq 2$. This gives the second product term in (4.25). Since, the access procedure going on for each access channel is independent of that for all other channels, therefore we can use the chain rule of probability for independent events, to get the first product term in (4.25), as shown by the following expression,

$$\Pr\Big(\big({}_jX_n\big)_{j=1}^{N_{\rm x}} = x_1^{N_{\rm x}}\Big|\big({}_jL_n\big)_{j=1}^{N_{\rm x}} = l_1^{N_{\rm x}}\Big) = \prod_{j=1}^{N_{\rm x}} \Pr\big({}_jX_n = x_j \,\Big|\, {}_jL_n = l_j\big), \tag{4.27}$$

$\forall x_1^{N_{\rm x}} \in \mathcal{S}_{\rm X}(x)$, $\forall x \in \Omega_{\rm X}$, $\forall l_1^{N_{\rm x}} \in \mathcal{S}_{\rm L}(l)$ and $l \geq 2$. Note that during an nth AP, for jth access control channel, if ${}_jL_n = 1$ then there will always be an access success and ${}_jX_n = 1$ with probability 1. Otherwise, if ${}_jL_n \geq 2$, collision occurs and the users go into the CRP mode, thus the access success for nth AP depends on the outcome of the CRP:

$$\Pr\big({}_jX_n = 1 \,\Big|\, {}_jL_n = l_j\big) = \begin{cases} 0 & \text{if } l_j = 0, \\ 1 & \text{if } l_j = 1, \\ \Psi(l_j) & \text{if } l_j \geq 2. \end{cases} \tag{4.28}$$

$\forall l_j \in \overline{\Omega}_{\rm L}$ and $\forall j \in \{0, 1, \ldots, N_{\rm x}\}$. Here, $\Psi(l_j)$ is the probability of access success for CRP. Now we discuss the evaluation of $\Psi(l_j)$. To this end, first let $\psi(l_j)$ be the probability that exactly one out of $l_j \geq 2$ contenders contends for an access slot during CRP. We also call it the *access success probability per access slot during CRP*, and it is given by:

$$\psi(l_j) = l_j\sigma(1-\sigma)^{l_j-1}, \tag{4.29}$$

$\forall l_j \in \overline{\Omega}_L$, $l_j \geq 2$, $j \in \{0, 1, \ldots, N_{\rm x}\}$. As a result, the probability of access success for the kth access slot during CRP is $\big[1-\psi(l_j)\big]^{k-1}\psi(l_j)$. Hence, the probability of access success for CRP, with at the most s access slots available per AP for CRP, is given by:

$$\Psi(l_j) = \sum_{k=1}^{s} \big[1-\psi(l_j)\big]^{k-1}\psi(l_j). \tag{4.30}$$

$\forall l_j \in \overline{\Omega}_{\mathrm{L}}$ and $\forall j \in \{0, 1, \ldots, N_{\mathrm{x}}\}$. Note that here we present $\Psi(l_j)$ for the assumed CRP, but any other protocol can be incorporated in our model by simply updating $\Psi(l_j)$ accordingly. This completes the evaluation of $\Pr\big({}_jX_n = 1 \,\big|\, {}_jL_n = l_j\big)$. Since ${}_jX_n \in \{0, 1\}$, thus $\Pr\big({}_jX_n = 0\big) = 1 - \Pr\big({}_jX_n = 1\big)$. This concludes the evaluation of the first product term in (4.25). The second term is already given by (4.26). Thus, this also concludes the evaluation of the summand in (4.23), using (4.25). We can now use (4.23) to evaluate the required conditional distribution $\Pr\big(X_n = x \big| L_n = l\big)$ using (4.21).

4.3.6 Optimization Problem - The Joint Channel Allocation

Our objective is to find the optimal number of access control and communication channels. The criterion for optimality is the minimization of the total loss rate of the system β. At the same time we also have two constraints. The first constraint is that there should always be at least one control and one communication channel in the system. According to this constraint we need to select the optimal N_{x} and N_{c} out of the set $\{1, \ldots, N-1\}$. The second constraint is that the total number of channels in the system remains constant. According to this constraint we have $N_{\mathrm{x}} + N_{\mathrm{c}} = N$. Moreover, we also need to select an optimal σ that minimizes the loss rate, for any given channel allocation. Thus, for a given N total number of channels, the optimal number of control channels N_{xo} is as follows:

$$N_{\mathrm{xo}} = \underset{N_{\mathrm{x}} \in \{1,\ldots,N-1\}}{\arg\min} \hat{\beta}(N_{\mathrm{x}}, N_{\mathrm{c}}), \text{ subject to: } N_{\mathrm{x}} + N_{\mathrm{c}} = N. \tag{4.31}$$

$$= \underset{N_{\mathrm{x}} \in \{1,\ldots,N-1\}}{\arg\min} \hat{\beta}(N_{\mathrm{x}}, N - N_{\mathrm{x}}). \tag{4.32}$$

Here, $\forall N_{\mathrm{x}} = N - N_{\mathrm{c}} \in \{1, \ldots, N-1\}$,

$$\hat{\beta}(N_{\mathrm{x}}, N_{\mathrm{c}}) = \min_{\sigma \in (0,1)} \beta(N_{\mathrm{x}}, N_{\mathrm{c}}, \sigma). \tag{4.33}$$

Note that here we have considered the variation in β with respect to N_{x}, N_{c} and σ, even though β depends on other model parameters as well which we assumed to be known for the stated optimization problem. According to the constraint, the optimal number of communication channels is

$N_{\text{co}} = N - N_{\text{xo}}$. Also note that this is a discrete optimization problem and does not have a closed form solution. Therefore, we numerically solve the problem by exhaustive search and analyze the results in Section 4.6. However, in a practical system, instead of finding the optimal channel allocation by exhaustive search, an already generated *channel allocation map* is stored in the system's memory as a lookup table which is used for an instantaneous channel allocation, whenever there is a change in traffic load. As discussed in detail in Section 4.6.4, the channel allocation map provides the optimal channel allocation for all possible values of the traffic parameters, namely, λ and ω. Note that while generating a channel allocation map, the optimal values of σ are also obtained which can also be stored in the form of a look table, for all values of λ and ω.

4.4 Traffic Aware Channel Allocation

For the given values of *system parameters* M, N and s, we can solve (4.31) and find the optimal number of control and communication channels, provided we know the values of the *traffic parameters* λ and ω. Recall from Section 4.2 that ω is a traffic parameter that is visible to the system, whereas λ is a parameter that is invisible. However, we can develop an estimator based on our system model, which uses the values of visible parameters, namely, X_n, Y_n and Z_n, from the available system data, to estimate the *invisible actual traffic load* λ. One such estimator is the *maximum likelihood (ML) estimator* [92], discussed in Section 4.4.1. Once we get an estimate of λ, say $\hat{\lambda}$, we can then allocate the control and communication channels optimally, for the known values of ω and the system parameters, as demonstrated by Figure 4.4.

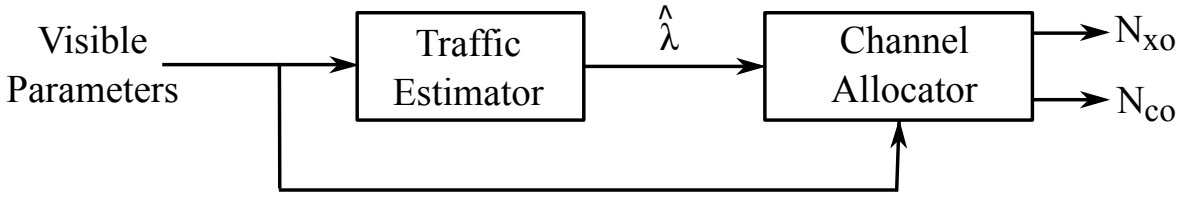


FIGURE 4.4: Traffic aware channel allocation.

4.4.1 Estimation of Invisible Actual Traffic

Here we present the ML estimator for the invisible actual traffic load λ. Firstly, we assume that λ is an unknown parameter whose value needs to be estimated. Now the ML estimate of λ, until time $k \in \mathbb{N}^+$, is given by:

$$\hat{\lambda}_k = \underset{a\in(0,1)}{\arg\max}\ \mathcal{L}(x_0^k, y_0^k, z_0^k, a). \tag{4.34}$$

Here $\mathcal{L}(.)$ is the likelihood of the visible parameters given a particular value of $\lambda = a$,

$$\begin{aligned}
&\mathcal{L}(x_0^k, y_0^k, z_0^k, a) \\
&= \Pr\big(X_0^k = x_0^k, Y_0^k = y_0^k, Z_0^k = z_0^k \big| \lambda = a\big) &(4.35)\\
&= \prod_{n=1}^{k} \Big(\Pr\big(Z_n = z_n \big| X_n = x_n, Y_n = y_n\big) \times \Pr(Y_n = y_n \big| Z_{n-1} = z_{n-1}) \Big) &(4.36)\\
&\quad \times \prod_{n=1}^{k} \Pr\big(X_n = x_n \big| Z_{n-1} = z_{n-1}, \lambda = a\big). &(4.37)
\end{aligned}$$

Here, in order to break up (4.35) into products, we used the chain rule of probability, and the characteristics of our model as described in Sections 4.2 and 4.3, i.e., $X_0 = Y_0 = Z_0 = 0$ with probability 1, Z_n is independent of all other random variables given X_n and Y_n, Y_n is independent of all other random variables given Z_{n-1}, and X_n is independent of all other random variables given Z_{n-1} and λ. Also, $X_0^k = (X_0, \ldots, X_k)$, Y_0^k, Z_0^k are $(k+1)$ dimensional vectors. Note that (4.36) is independent of λ and thus does not play a role in the ML estimator. Rather, only (4.37) depends on λ and therefore:

$$\hat{\lambda}_k = \underset{a\in(0,1)}{\arg\max} \prod_{n=1}^{k} \Pr\big(X_n = x_n \big| Z_{n-1} = z_{n-1}, \lambda = a\big). \tag{4.38}$$

Hence, the ML estimator only requires the conditional distribution of the access state and it can be evaluated using our model, for each $n \in \mathbb{N}^+$, as detailed in Section 4.3.4.

4.4.2 Learning the Optimal Channel Allocation

Recall from Section 4.3.4 that for a given value of $\lambda = a$, the evaluation of the access state distribution $\Pr\big(X_n = x_n \big| Z_{n-1} = z_{n-1}, \lambda = a\big)$ requires values of the model parameters, namely, M,

N_{x}, N_{c}, s, σ and ω. Since this distribution is needed to estimate λ, as shown in (4.38), therefore we need to know the values of N_{x}, N_{c} and σ to get the ML estimate of λ, i.e., $\hat{\lambda}$, for the given values of M, N, s and ω. Hence, we can start with arbitrary values of N_{x}, N_{c} and σ to get an initial estimate of λ, using our proposed model and solving (4.38). Later, we can use this estimate of the invisible traffic load to update the optimal values of N_{x}, N_{c} and σ by using our proposed model and solving (4.31) and (4.33). Recall from Section 4.3.6 that in practical systems, instead of solving (4.31) and (4.33) every time the load varies, we use the already generated look tables stored in the system's memory to get the optimal values of N_{x}, N_{c} and σ, for the given values of λ and ω. This basic idea can help develop an iterative algorithm for the traffic aware joint allocation of control and communication channels in the system. In this study, we focus on the development and analysis of the model required for such an algorithm, while a more comprehensive study on the development of the algorithm is left for future work.

4.5 Reported Loss Rate

In contrast to the actual total loss rate β, defined in Section 4.3, we now define the *reported loss rate* of the system based on the visible parameters. We define it as the *fraction of visible-calls that are blocked* during an SP in the long run. Here, by *visible-calls* we mean those calls that successfully get access to the system and are thus recorded in the practical system data. Also, the blocked calls that are visible to the system are only those which are blocked due to congestion. Therefore, similar to (4.7) and (4.9), we can evaluate the reported loss rate as follows, for $\sum_{\forall z_1 \in \Omega_{\mathrm{Z}}} \mathrm{E}\big[X_n \big| Z_{n-1} = z_1\big].\pi_{z_1} \neq 0$:

$$\tilde{\beta} = \lim_{k\to\infty} \frac{\sum_{n=1}^{k} G_n}{\sum_{n=1}^{k} X_n} \tag{4.39}$$

$$= \frac{\sum_{\forall z_1 \in \Omega_{\mathrm{Z}}} \mathrm{E}\big[G_n \big| Z_{n-1} = z_1\big].\pi_{z_1}}{\sum_{\forall z_1 \in \Omega_{\mathrm{Z}}} \mathrm{E}\big[X_n \big| Z_{n-1} = z_1\big].\pi_{z_1}}. \tag{4.40}$$

Recall from Section 4.3.1 that the numerator in (4.40) is the expected number of calls blocked due to congestion, while the denominator is the expected number of calls successfully getting access to the system, during an SP in the long run. Also note that from (4.7), (4.9) and (4.40), and using

the fact that $Q_n = L_n - X_n$, we can easily derive:

$$\tilde{\beta} = \frac{\beta_{\mathrm{G}}}{1 - \beta_{\mathrm{Q}}}. \tag{4.41}$$

We call $\tilde{\beta}$ as the reported loss rate because it is usually reported by the practical system administrators as a system performance metric. The administrators evaluate this loss rate using the data of call records stored by the practical system. In Section 4.6.2 we demonstrate that the reported loss rate, that is used by the practical system administrators, is a misleading performance metric, as compared to the actual loss rate that can be estimated using our model as explained in Section 4.5.1.

4.5.1 Estimating the Actual Loss Rate

Recall that the actual loss rate can be evaluated with the help of our model, as explained in Section 4.3. However, as discussed in the same section, this requires the knowledge of the number of contenders L_n and the collision loss $Q_n = L_n - X_n$. Note that both L_n and Q_n are invisible system parameters and depend on the invisible parameter λ, as discussed in Sections 4.2 and 4.3. In Section 4.4, we explain how we can use our model to estimate λ, which can then be used to evaluate the expected values of L_n and Q_n, and finally enable us to evaluate the actual loss rate for the system.

4.6 Results and Discussion

The problem under consideration is represented by (4.31). Recall that it is a discrete optimization problem and does not have a closed form solution. Thus, in this section, we numerically solve this optimization problem and analyze the results. Also, we assume that λ is known. Note that we only consider λ and ω lying in open intervals $(0, 1)$. We ignore the values, 0 and 1, since they are not practical cases. Also note that in order to find the optimal channel allocation, we first minimize the total loss rate over σ, for each possible channel allocation, and later minimize the same over all possible allocations. Also, in this section we consider $M = 10$ and $N = M/2 = 5$, which is a frequent practical scenario observed in public safety radio networks [64]. Note that in finite

source systems, e.g., public safety trunked radio systems, M represents the number of talk groups, instead of the actual number of radio units, since a single talk group can have many radio units as its members. Indeed, from the meta data collected on *Louisiana Wireless Information Network (LWIN)*, $M = 10$ groups could correspond to as many as hundreds of radio units [64].

4.6.1 Simulation Results:

Recall that the objective function of the optimization problem under consideration is the total loss rate of the system. Figure 4.5 shows that the value of the total loss rate, which is estimated via simulation, approaches the value which is numerically evaluated using our proposed framework, for a large number of *regenerative* simulation cycles. This demonstrates the validity of the numerical results presented in this section. Note that the total loss rate which is numerically evaluated using our proposed framework is $\beta = \frac{\sum_{\forall z_1 \in \Omega_{\mathrm{Z}}} \mathrm{E}\left[Q_n + G_n \middle| Z_{n-1}=z_1\right].\pi_{z_1}}{\sum_{\forall z_1 \in \Omega_{\mathrm{Z}}} \mathrm{E}\left[L_n \middle| Z_{n-1}=z_1\right].\pi_{z_1}}$, whereas the one estimated at the kth simulation time is $\frac{\sum_{n=1}^{k} Q_n + G_n}{\sum_{n=1}^{k} L_n}$. Also note that we use the *discrete event Monte Carlo simulation* with the *regenerative method* [41].

4.6.2 Reported versus Actual Loss Rate - Misleadingness of Reported Loss

Both the actual and the reported loss rates of the system monotonically increase with increase in call arrival traffic λ, as shown in Figure 4.6. Recall that the actual loss rate β incorporates both the collision and congestion losses, whereas the reported loss rate $\tilde{\beta}$ only incorporates the congestion loss. Due to this reason, a significant difference between both the performance measures is evident in Figure 4.6. This clearly shows that the reported loss rate under estimates the system loss, and therefore, it is quite misleading to have it as a performance metric.

4.6.3 Existence of N_{xo} - Collision and Congestion Trade-off

Increasing the number of control channels N_{x} in the system, while keeping the total number of channels constant, decreases the number of communication channels N_{c}, and thus has a two-folds effect on the total system loss. On one hand an increase in N_{x} decreases the collision loss by increasing the expected number of calls successfully getting access to the system. On the other hand, due to the same reason, the traffic load for the reduced number of communication channels

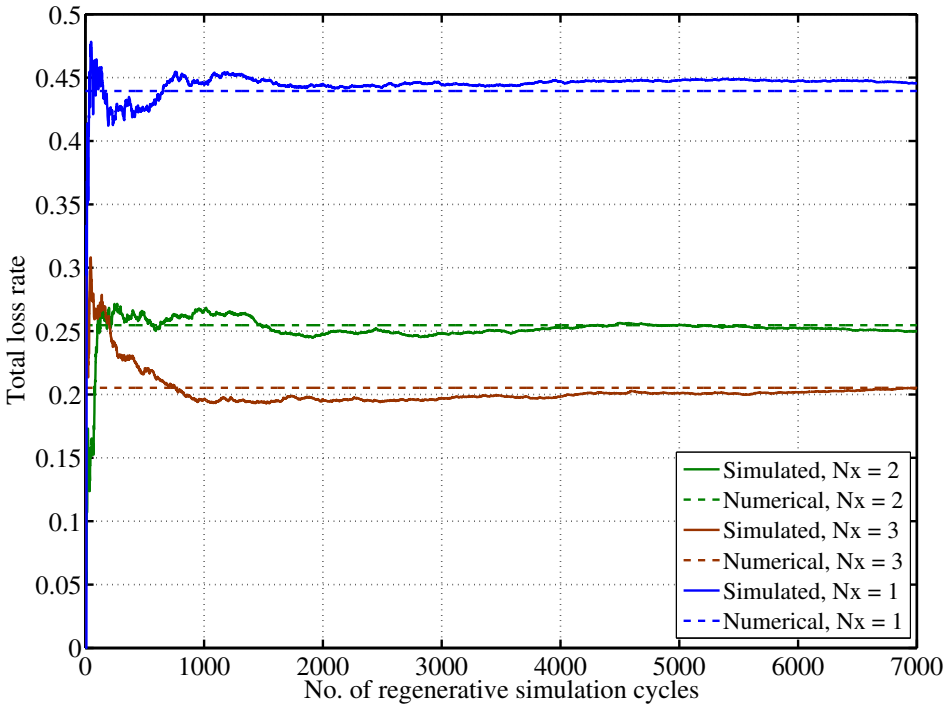


FIGURE 4.5: Simulation results ($N = 5$, $M = 10$, $s = 4$, $\lambda = \omega = \sigma = 0.1$).

increases, which increases the congestion loss. Thus, with an increase in N_{x} there exists a trade-off between collision and congestion losses. In Figure 4.7, this trade-off is clearly evident. For $N_{\mathrm{x}} < 2$, the increase in congestion loss is dominated by the decrease in collision loss, thereby decreasing the total loss rate. At $N_{\mathrm{x}} = 2$, the minimum total loss rate is achieved for the selected system and traffic parameters. Now, if N_{x} is further increased beyond this point, i.e., for $N_{\mathrm{x}} > 2$, the decrease in collision loss is dominated by the increase in congestion loss, thereby decreasing the total loss rate. Thus there exists an optimal number of control channels for the system, e.g., $N_{\mathrm{xo}} = 2$ in Figure 4.7, which corresponds to an optimal trade-off between collision and congestion performance of the system, and minimizes the total loss rate. Also note that in order to find the optimal channel allocation, we first minimize the total loss rate over σ, for each possible channel allocation, and later minimize the same over all possible allocations, as shown in (4.33). Therefore, In Figure 4.7, we have also plotted the corresponding optimal values of σ.

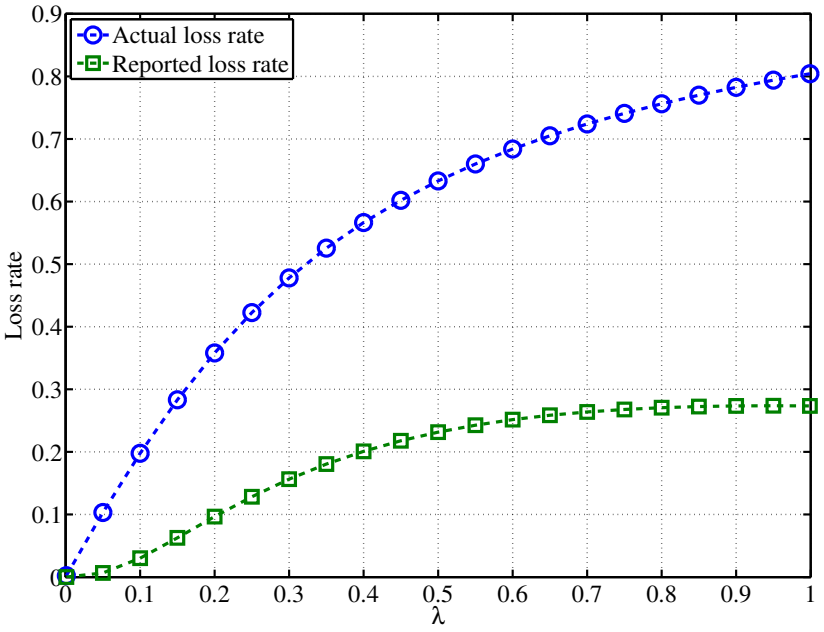


FIGURE 4.6: Reported versus actual loss rate ($N_{\mathrm{x}} = 2$, $N_{\mathrm{c}} = 3$, $M = 10$, $\omega = 0.5$, $s = 10$, $\sigma = 0.4$).

4.6.4 Channel Allocation Map

We define the *channel allocation map* as a color map that is used to present the optimal number of control channels N_{xo}, for all possible values of the traffic parameters, namely, λ and ω, for the given values of system parameters, namely, M, N and s. Also note that the optimal number of communication channels are simply $N_{\mathrm{co}} = N - N_{\mathrm{xo}}$. One such map is shown in Figure 4.8(a), for $N = 5$ and $M = 10$, for a system without CRP, i.e., $s = 0$. In this map, the color indicates the value of N_{xo} for the system, for a certain (λ, ω)-pair. Now we shall analyze the variation of N_{xo} with respect to λ and ω in the map shown in Figure 4.8(a).

For any fixed λ, the optimal number of control channels N_{xo} decreases with increase in ω, in Figure 4.8(a). This is because, for a fixed λ, increase in ω means an increase in expected call duration which increases the congestion loss. It also decreases the expected number of idle users per SP, thereby decreasing the expected number of contenders and the collision loss. This increase in congestion and decrease in collision requires an increase in the number of communication

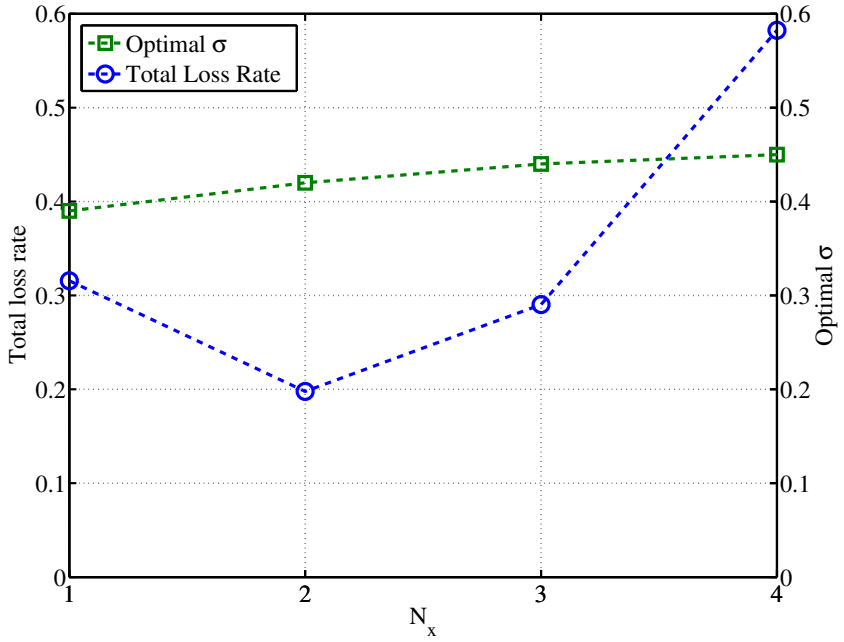
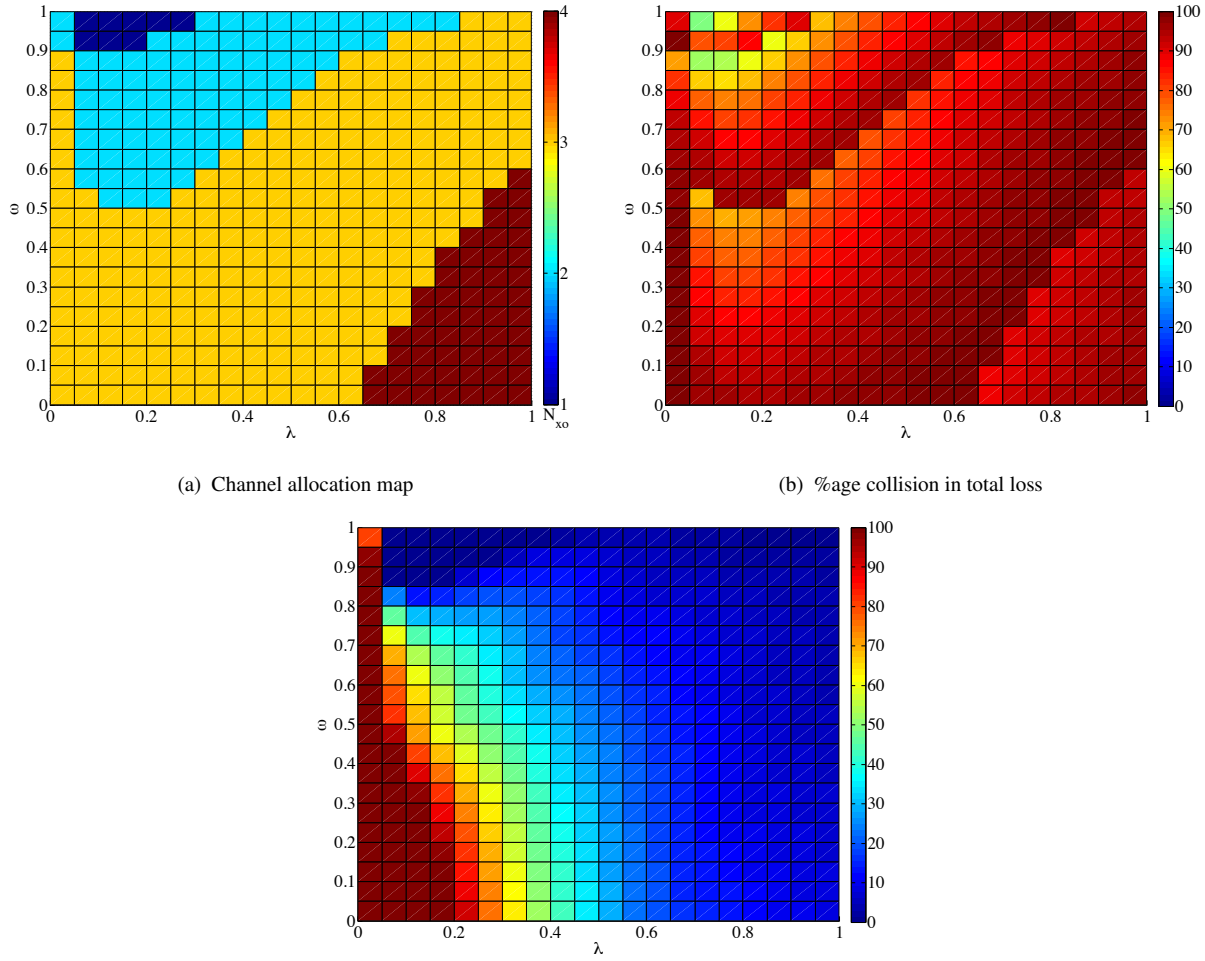


FIGURE 4.7: Existence of N_{xo} ($N = 5$, $M = 10$, $\lambda = 0.1$, $\omega = 0.5$, s = 10).

channels. Thus, the optimal number of control channels N_{xo} decreases with increase in ω, for a fixed λ. Note that this coupling between collision and congestion losses is captured by our finite source modeling of the system, as detailed in Sections 4.1 and 4.2.

For the variation in N_{xo} with respect to λ, for a fixed ω, we need to divide the map in Figure 4.8(a) into two main regions, namely, low-ω and high-ω regions. Consider the low-ω region in Figure 4.8(a), $\omega \in (0, 0.5)$. Recall that an increase in λ increases the overall call traffic in the system, which increases both collision and congestion losses. In the low-ω region, due to the low value of ω, the increase in congestion loss with increase in λ is dominated by the increase in collision loss. Therefore, in this region, the optimal number of control channels N_{xo} either increases or remains constant with increase in λ. This traffic scenario is similar to the *small data packet transmission* problem [93], for modern cellular systems.

Next, consider the high-ω region in Figure 4.8(a), $\omega \in (0.5, 1)$. We divide this region into two sub-regions, namely, high-ω-low-λ and high-ω-high-λ sub-regions. Consider the high-ω-low-λ sub-region, which is the region with $\omega \in [0.5, 1)$ and $\lambda \in (0, 0.2)$. In this sub-region, due to the



(a) Channel allocation map

(b) %age collision in total loss

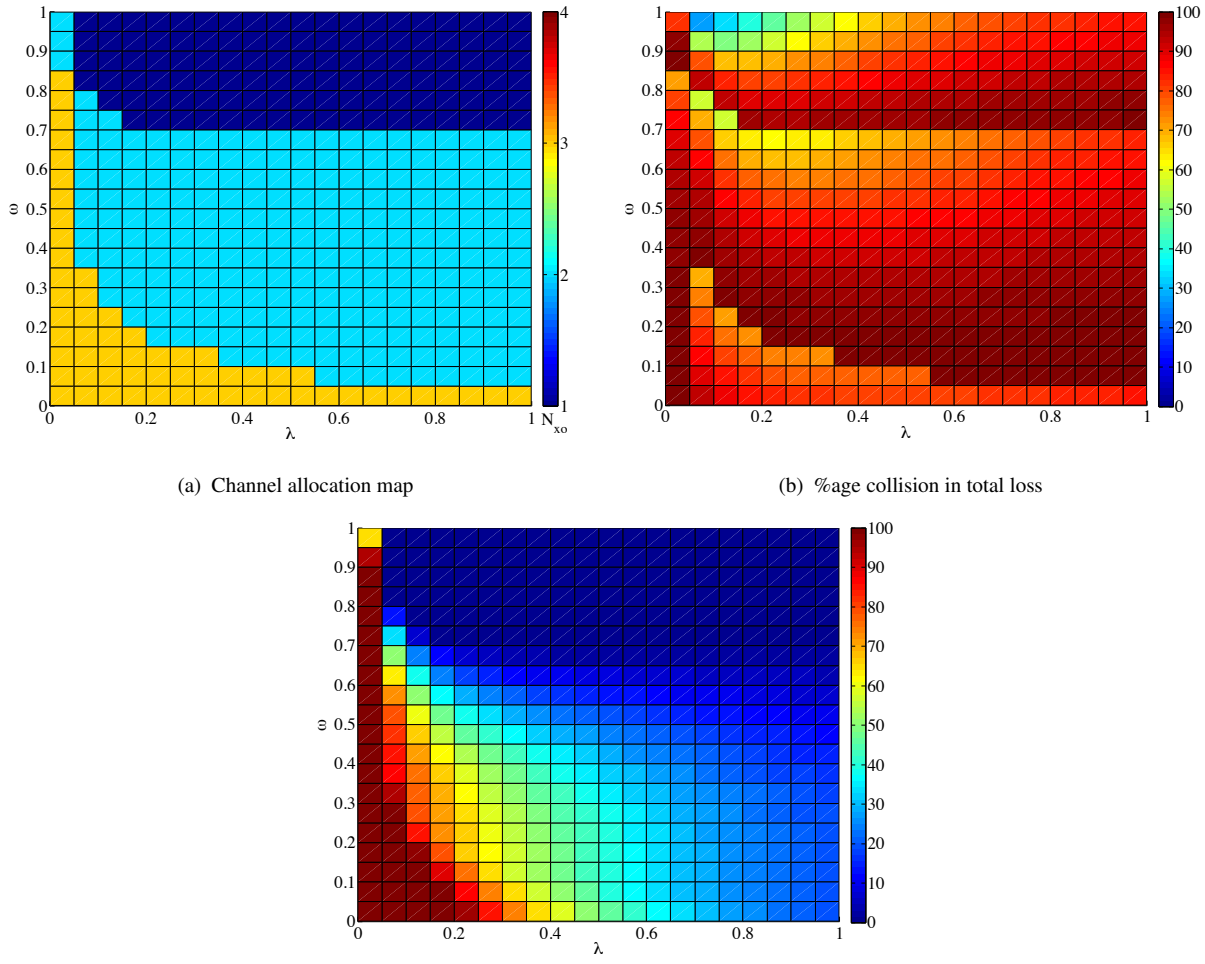
(c) %age loss increment for N_x=1 w.r.t optimal

FIGURE 4.8: Numerical results without CRP ($s = 0$, $N = 5$, $M = 10$).

high value of ω and low value of λ, the increase in collision loss is dominated by the increase in congestion loss, as λ increases within this sub-region. Therefore, in this region, the optimal number of control channels N_{xo} either decreases or remains constant with increase in λ.

Finally we consider the high-ω-high-λ sub-region in Figure 4.8(a), which is the region with $\omega \in [0.5, 1)$ and $\lambda \in [0.2, 1)$. In this sub-region, the value of λ is so high that the increase in congestion loss is dominated by the increase in collision loss, as λ increases within this sub-region, even though ω is high as well. Therefore, in this region, the optimal number of control channels N_{xo} either increases or remains constant with increase in λ.

Channel allocation map with CRP: Figure 4.8(a) shows that without CRP, the optimal number of control channels is $N_{\text{xo}} \geq 3$ for a significant part of the traffic region, for the given system parameters, $N = 5$ and $M = 10$. Also, without CRP, a very small part of the traffic region has a single control channel as an optimal allocation. On the other hand, Figure 4.9(a) shows that for the system with CRP the optimal number of control channels is $N_{\text{xo}} \leq 2$ for a significant part of the traffic region. This decrease in optimal number of control channels, in the systems with CRP, is due to the fact that the collision loss decreases due to use of CRP, therefore the system requires a lesser number of control channels to achieve the optimal net performance in terms of the minimum total loss rate.



(a) Channel allocation map

(b) %age collision in total loss

(c) %age loss increment for N_{x}=1 w.r.t optimal

FIGURE 4.9: Numerical results with CRP ($s = 10$, $N = 5$, $M = 10$).

4.6.5 Time versus Frequency Resources in Control Layer

In our model, we represent the number of access slots per AP for CRP by s. It is the maximum number of access retrials that are allowed per AP in the system. It is also the measure of time resources spent for CRP at the control layer of the system. So far, we have seen that with CRP we can reduce the required number of control channels, by spending more time resources at the control layer with $s > 0$. But time resources spent for CRP at the control layer, or s, cannot be increased infinitely, because of the limitations imposed by the control layer design and the required signaling rate. Thus, s is a bounded resource and cannot be increased to an infinitely large value. Moreover, for finite values of s, we may need more than one frequency control channel, depending on the traffic parameters, as shown by our results in Figure 4.9(a).

4.6.6 Significance of Collision Loss

Figures 4.8(b) and 4.9(b) show the percentage of collision loss in total loss, for the optimal channel allocations shown in respective Figures 4.8(a) and 4.9(a), for different values of λ and ω. Figures 4.8(b) and 4.9(b) show that for $N = 5$ and $M = 10$, more than 40% of the total loss is due to collision. In fact, for most of the traffic region, collision loss is more than 70% of the total loss. This indicates that even for the optimal channel allocation, collision loss is a significant part of the total loss, and hence cannot be ignored. Now we shall explain why the percentage of collision is so high. Firstly consider the case when collision is very high, e.g., for very high λ, and thus the expected number of calls successfully getting access to the system is very low. This results in a very low congestion loss. Therefore, in this scenario, collision loss becomes a significant part of the total loss. On the other hand, for a relatively lower collision loss, one can reason that the collision loss per SP can go as high as the total number of users in the system, whereas the congestion loss per SP can only go as high as the maximum number of users that can successfully get system access, which is equal to the number of control channels. Since the number of users is usually much greater than the number of channels, therefore, most of the time,

even if the absolute magnitudes of the collision and congestion loss rates are low, the collision loss can proportionately be a significant part of the total loss, as compared to congestion.

4.6.7 Comparison with Single Control Channel System

Figures 4.8(c) and 4.9(c) show the percentage increase in total loss rate when a single control channel is used instead of the optimal allocation, for $N = 5$ and $M = 10$. Let Δ_{p} represent this percentage increase in loss rate, then we define it mathematically as:

$$\Delta_{\mathrm{p}} = \frac{\beta(N_{\mathrm{x}} = 1) - \beta(N_{\mathrm{x}} = N_{\mathrm{xo}})}{\beta(N_{\mathrm{x}} = N_{\mathrm{xo}})} \times 100. \tag{4.42}$$

Figures 4.8(c) and 4.9(c) show that for most of the traffic region $\Delta_{\mathrm{p}} \geq 20\%$, for $N = 5$ and $M = 10$. Hence for these values of system parameters, most of the time, using a single control channel, instead of the optimal channel allocation, results in at least 20% more loss rate, as compared to the optimal allocation. This shows the significance of the optimal channel allocation as compared to the conventional single control channel system.

4.7 Conclusion

We proposed a novel model for finite source multiple access systems which jointly models the access-control and communication layers. We then formulated the optimization problem to *jointly* find the optimal number of control and communication channels that minimizes the actual total loss rate, for a given number of channels available in total. Due to its complexity, we solve the problem numerically. Firstly, we demonstrated the existence of the optimal solution using numerical results. Later, we introduced the concept of a *channel allocation map* to represent the optimal channel allocation for all possible values of the traffic parameters. As a further contribution, we used our model to quantify the reported system loss rate and elaborated its misleadingness as compared to the actual loss rate. We also demonstrated a mechanism, based on our proposed model, which can be used to estimate the invisible actual traffic load, required for deciding the *traffic-aware* optimal channel allocation, and also for estimating the actual loss rate, for a practical system. Moreover, we provided guidelines for developing an algorithm for the traffic aware

allocation of control and communication channels, based on our proposed model and its capability to estimate the actual invisible traffic load. Using numerical results, we demonstrated that the optimal channel allocation provides a significant improvement in performance as compared to the conventional strategy of using a single control channel. Our results also showed that even though we spend time resources at the control layer to alleviate collision, we may still need more than one frequency channel for access-control, depending on the traffic parameters.

Chapter 5
Segmentation of Talk Group's Call Activity

The spectrum allocation techniques, e.g., the ones discussed in Part-I of this study, requires the knowledge of the call traffic parameters and the priority levels of the users in the system. For practical systems, these required pieces of information are extracted from the call records meta-data. A key fact that should be considered while analyzing the call records is that the call arrival traffic and the users priority levels change with a change in events on the ground. This is so because a change in events on the ground affects the communication behavior of the talk groups in the system, which affects the call arrival traffic and the priority levels of the users. For example, a change from a normal scenario to an emergency situation, which causes a change in the talk groups' activities of a PSCS. Thus, the first and the foremost step in analyzing the call records data for a given talk group, for extracting the call traffic information, is to segment the data into time intervals of homogeneous or stationary communication behavior of the group. In this chapter we develop an algorithm for such a segmentation of the data of a practical PSCS.

In Section 5.6.1, we propose a way to quantify a talk group's activity as a discrete sequence of symbols. Here a symbol, corresponding to a call, represents the calling radio-unit in the talk group. Note that we extract this talk group activity from the available data of call records. Recall that in practice, a talk group's communication behavior remains consistent during a certain event happening on the ground. However, with a change in event, the talk group's behavior also changes. Thus, a talk group's activity, extracted as a discrete sequence, from the whole day data of call records, is not stationary. However, it may consist of many stationary segments depending on different events that occurred on the ground. Our primary goal is to segment a given whole day activity of a talk group into stationary segments, each corresponding to a distinct event, and also quantify the behavior of the talk group within each segment or event. Note that we quantify the

behavior of a talk group in the form of a *context tree* which represents a *variable memory Markov process* of discrete symbols. The manuscript of a research paper based on this project is under preparation [94].

5.1 Context Tree Model for Variable Length Markov Chain

Here we briefly describe the context tree based model for a variable length Markov chain (VLMC). Further details can be found in [95]. Let $\boldsymbol{x} = x_0^n = (x_0, \ldots, x_n)$ be the given sequence of symbols with length $l(\boldsymbol{x}) = n + 1$, which is to be segmented, and $\mathcal{A}(\boldsymbol{x})$ be the set of finite alphabets in $\boldsymbol{x}$. We also represent an indexed list, i.e., a vector, of the symbols in $\mathcal{A}(\boldsymbol{x})$ as $\overline{a}(\boldsymbol{x})$. Also note that the given sequence $\boldsymbol{x} = x_0^n = (x_0, \ldots, x_n)$ represents a realization of a stationary ergodic stochastic process, $\boldsymbol{X} = X_0^n = (X_0, \ldots, X_n)$, [95]. The concatenation of sequences $\boldsymbol{s}$ and $\boldsymbol{u}$ is $\boldsymbol{su}$. We say that a sequence $\boldsymbol{u}$ is a *suffix* of $\boldsymbol{w}$, denoted by $\boldsymbol{u} \preceq \boldsymbol{w}$, if there exists an $\boldsymbol{s}$ such that $\boldsymbol{w} = \boldsymbol{su}$. However, $\boldsymbol{u}$ is a *proper suffix* of $\boldsymbol{w}$, i.e., $\boldsymbol{u} \prec \boldsymbol{w}$, if $\boldsymbol{u} \neq \boldsymbol{w}$. We also denote an empty sequence as λ with $l(\lambda) = 0$. A VLMC that generates the sequence $\boldsymbol{x}$ can be modeled as a *context tree* [95]. This context tree is represented by a list $\overline{\tau}$ of sequences called *contexts*. These contexts are the leaf nodes of the context tree. A set of all the nodes of the context tree can be represented as $\mathcal{T}(\overline{\tau})$. These nodes are the finite suffixes of all the contexts. Each context $\boldsymbol{w} = w_i^j$ in $\overline{\tau}$ is visualized as a path from the leaf $\boldsymbol{w}$ to the root λ, consisting of $l(\boldsymbol{w})$ edges, each labeled by the symbols $w_i, \ldots, w_j$. Corresponding to each context $\boldsymbol{w}$ in $\overline{\tau}$, there is an indexed list of transition probabilities $\overline{\pi}(\boldsymbol{w})$ such that its jth element is $\pi_j(\boldsymbol{w}) = \Pr(X_i = a_j \big| X_{i-l(\boldsymbol{w})}^{i-1} = \boldsymbol{w})$, for all i such that $l(\boldsymbol{w}) \leq i \leq l(\boldsymbol{x})$. Here $0 \leq j < |\overline{a}|$. We also define depth of context tree as $d(\overline{\tau}) = \max\{l(\boldsymbol{w}), \boldsymbol{w} \in \overline{\tau}\}$.

5.1.1 BIC Estimation of Context Tree Model

An algorithm is proposed in [95] that estimates a context tree for a given sequence $\boldsymbol{x}$ according to the *Bayesian Information Criterion (BIC)*. Let $N_{\boldsymbol{w}}(\boldsymbol{x})$ denote the number of occurrences of the sequence $\boldsymbol{w}$ in the given sequence $\boldsymbol{x}$, such that $l(\boldsymbol{w}) \leq D$. Here, the given parameter D is the

maximum possible depth of the context tree [95]. Mathematically, $N_{\boldsymbol{w}}(\boldsymbol{x})$ can be defined as,

$$N_{\boldsymbol{w}}(\boldsymbol{x}) = \left| \left\{ i : l(\boldsymbol{w}) \leq i \leq l(\boldsymbol{x}), x_{i-l(\boldsymbol{w})}^{i-1} = \boldsymbol{w} \right\} \right|. \tag{5.1}$$

For a given sequence $\boldsymbol{x}$, a *feasible tree* [95] is any tree $\overline{\tau}$ of depth $d(\overline{\tau}) \leq D$, such that $N_{\boldsymbol{w}}(\boldsymbol{x}) \geq 1$, $\forall \boldsymbol{w} \in \overline{\tau}$, and each string $\boldsymbol{s}$ with $N_{\boldsymbol{s}}(\boldsymbol{x}) \geq 1$ is either a suffix of some $\boldsymbol{w} \in \overline{\tau}$ or has a suffix $\boldsymbol{w} \in \overline{\tau}$. Let $\mathcal{F}_D(\boldsymbol{x})$ denote the set of all feasible trees of depth $d(\overline{\tau}) \leq D$ for a given sequence $\boldsymbol{x}$. Also, the maximum likelihood (ML) estimate of $\pi_j(\boldsymbol{w}) = \Pr(X_i = a_j \big| X_{i-l(\boldsymbol{w})}^{i-1} = \boldsymbol{w})$, for all i such that $l(\boldsymbol{w}) \leq i \leq l(\boldsymbol{x})$, is given by [95]:

$$\hat{\pi}_j(\boldsymbol{x}, \boldsymbol{w}) = \frac{N_{\boldsymbol{w}a_j}(\boldsymbol{x})}{\sum_{a_j \in \overline{a}(\boldsymbol{x})} N_{\boldsymbol{w}a_j}(\boldsymbol{x})}, \tag{5.2}$$

for all $0 \leq j < |\overline{a}(\boldsymbol{x})|$, and the maximum likelihood of the given sequence $\boldsymbol{x}$ with a feasible tree $\overline{\tau}$ is given by [95]:

$$\hat{\mathcal{L}}\big(\boldsymbol{x}, \overline{\tau}\big) = \prod_{\boldsymbol{w} \in \overline{\tau}} L_{\boldsymbol{w}}(\boldsymbol{x}), \tag{5.3}$$

$$L_{\boldsymbol{w}}(\boldsymbol{x}) = \prod_{j=0}^{|\overline{a}(\boldsymbol{x})|-1} \hat{\pi}_j(\boldsymbol{x}, \boldsymbol{w})^{N_{\boldsymbol{w}a_j}(\boldsymbol{x})}. \tag{5.4}$$

We also define $\mathrm{df}(\overline{\tau}, \boldsymbol{x})$ as the degrees of freedom of the context tree $\overline{\tau}$ for a given sequence $\boldsymbol{x}$, and it is given by [95]:

$$\mathrm{df}(\overline{\tau}, \boldsymbol{x}) = \sum_{\boldsymbol{w} \in \overline{\tau}} \nu_{\boldsymbol{w}}(\boldsymbol{x}) = |\overline{\tau}|.(|\mathcal{A}(\boldsymbol{x})| - 1). \tag{5.5}$$

Here $|\overline{\tau}|$ represents the number of contexts or leaf nodes in tree $\overline{\tau}$, and $\nu_{\boldsymbol{w}}(\boldsymbol{x}) = |\mathcal{A}(\boldsymbol{x})| - 1$ represents the degrees of freedom for a context $\boldsymbol{w}$ and it depends on the number of possible transitions from context $\boldsymbol{w}$ to symbols in $\mathcal{A}(\boldsymbol{x})$. The BIC estimate for the context tree is given by [95]:

$$\hat{\overline{\tau}}_c(\boldsymbol{x}) = \underset{\overline{\tau} \in \mathcal{F}_D(\boldsymbol{x})}{\arg\min}\, B_{\mathrm{S}}(\boldsymbol{x}, \overline{\tau}, c). \tag{5.6}$$

Here $B_{\mathrm{S}}(\boldsymbol{x}, \overline{\tau}, c)$ is the BIC score for the sequence $\boldsymbol{x}$, given a feasible context tree $\overline{\tau}$ and the *tree penalty constant* c.

$$B_{\mathrm{S}}(\boldsymbol{x}, \overline{\tau}, c) = -\log \hat{\mathcal{L}}\big(\boldsymbol{x}, \overline{\tau}\big) + c\,.\mathrm{df}(\overline{\tau}, \boldsymbol{x})\,.\log l(\boldsymbol{x}). \tag{5.7}$$

5.1.2 Algorithm for Context Tree Estimation

For a given sequence of symbols $\boldsymbol{x}$, tree penalty constant c, and maximum context tree depth D, a context tree satisfying the BIC criterion can be estimated by the procedure in Figure 5.1, proposed in [95]. Recall that the proper suffix relation '$\prec$', and the list of estimated transition

Estimate_Context_Tree(c, $\boldsymbol{x}$, D):

- Build a complete tree $\mathcal{T}$ (set of all nodes) with depth D.
- Recursively evaluate the following:

  Values: $V_{\boldsymbol{w}}^{c}(\boldsymbol{x})$, $\forall \boldsymbol{w} \in \mathcal{T}$.
  Indicators: $\delta_{\boldsymbol{w}}^{c}(\boldsymbol{x})$, $\forall \boldsymbol{w} \in \mathcal{T}$.
- $\hat{\overline{\tau}} = [\,], \hat{\overline{p}} = [\,]$.
- For $\boldsymbol{w} \in \mathcal{T}$:

  If $\delta_{\boldsymbol{v}}^{c}(\boldsymbol{x}) = 1$ for all $\boldsymbol{v} \prec \boldsymbol{w}$, and $\delta_{\boldsymbol{w}}^{c}(\boldsymbol{x}) = 0$:
  Append $\boldsymbol{w}$ to $\hat{\overline{\tau}}$, and $\overline{\hat{\pi}}(\boldsymbol{x}, \boldsymbol{w})$ to $\hat{\overline{p}}$.
- Return $\hat{\overline{\tau}}$, $\hat{\overline{p}}$.

FIGURE 5.1: The context tree estimation procedure.

probabilities $\overline{\hat{\pi}}(\boldsymbol{x}, \boldsymbol{w})$ for a node $\boldsymbol{w}$ in the context tree are defined and explained in Section 5.1. The values and indicators mentioned in the Estimate_Context_Tree procedure are given by the following equations, as mentioned in [95]:

$$V_{\boldsymbol{w}}^{c}(\boldsymbol{x}) = \begin{cases} \max\left\{\hat{V}_{\boldsymbol{w}}^{c}(\boldsymbol{x}), \tilde{V}_{\boldsymbol{w}}^{c}(\boldsymbol{x})\right\}, & \text{if } 0 \leq l(\boldsymbol{w}) < D, \\ \tilde{V}_{\boldsymbol{w}}^{c}(\boldsymbol{x}), & \text{if } l(\boldsymbol{w}) = D. \end{cases} \tag{5.8}$$

$$\delta_{\boldsymbol{w}}^{c}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } 0 \leq l(\boldsymbol{w}) < D \text{ and } \hat{V}_{\boldsymbol{w}}^{c}(\boldsymbol{x}) > \tilde{V}_{\boldsymbol{w}}^{c}(\boldsymbol{x}), \\ 0, & \text{if } 0 \leq l(\boldsymbol{w}) < D \text{ and } \hat{V}_{\boldsymbol{w}}^{c}(\boldsymbol{x}) \leq \tilde{V}_{\boldsymbol{w}}^{c}(\boldsymbol{x}), \\ 0, & \text{if } l(\boldsymbol{w}) = D. \end{cases} \tag{5.9}$$

Here $\hat{V}^c_{\boldsymbol{w}}(\boldsymbol{x})$ and $\tilde{V}^c_{\boldsymbol{w}}(\boldsymbol{x})$ are given by,

$$\hat{V}^c_{\boldsymbol{w}}(\boldsymbol{x}) = \sum_{a \in \mathcal{A}(\boldsymbol{x}): N_{a\boldsymbol{w}}(\boldsymbol{x}) \geq 1} V^c_{a\boldsymbol{w}}(\boldsymbol{x}). \tag{5.10}$$

$$\tilde{V}^c_{\boldsymbol{w}}(\boldsymbol{x}) = \log L_{\boldsymbol{w}}(\boldsymbol{x}) - c.\nu_{\boldsymbol{w}}(\boldsymbol{x}) \log l(\boldsymbol{x}). \tag{5.11}$$

A more comprehensive detail on this algorithm is provided in [95].

5.2 Context Tree based Sequence Segmentation

We represent a particular segmentation of the given data sequence $\boldsymbol{x}$ as a list $\bar{i}$ of segment end indices. Let $\mathcal{H}_{\mathrm{S}}$ be the set of all possible segmentations of $\boldsymbol{x}$. The kth segment of $\boldsymbol{x}$ is thus $x_{i_{k-1}+1}^{i_k}$, and its minimum BIC score for a given tree penalty constant c and the maximum tree depth D is given by

$$\beta_{\mathrm{S}}(x_{i_{k-1}+1}^{i_k}, c, D) = \min_{\overline{\tau} \in \mathcal{F}_D(x_{i_{k-1}+1}^{i_k})} B_{\mathrm{S}}(x_{i_{k-1}+1}^{i_k}, \overline{\tau}, c) \tag{5.12}$$

$$= B_{\mathrm{S}}\big(x_{i_{k-1}+1}^{i_k}, \hat{\overline{\tau}}_c(x_{i_{k-1}+1}^{i_k}), c\big) \tag{5.13}$$

Note that $i_{-1} = -1$. Thus, we assign a BIC score or a cost $\beta_{\mathrm{S}}(x_{i_{k-1}+1}^{i_k}, c, D)$ to the kth segment. We evaluate this cost by fixing c and then estimating the context tree for the given segment using the algorithm presented in Section 5.1.1. The BIC score of this tree is $\beta_{\mathrm{S}}(x_{i_{k-1}+1}^{i_k}, c, D)$. Now, for a fixed c, the BIC score of the complete data sequence $\boldsymbol{x}$ for a segmentation $\bar{i}$ is given by [96],

$$\beta_{\mathrm{D}}(\bar{i}, c) = \sum_{k=0}^{|\bar{i}|-1} \beta_{\mathrm{S}}(x_{i_{k-1}+1}^{i_k}, c, D) + \frac{|\bar{i}| - 1}{2}. \log l(\boldsymbol{x}). \tag{5.14}$$

In this study we also call this score as the *cost of segmentation* $\bar{i}$, for a given c, and denote it as $C(c, \bar{i}) = \beta_{\mathrm{D}}(\bar{i}, c)$. The first term in (5.14) is the sum of the tree scores for all the segments, whereas the second term is the segmentation penalty, as proposed in [96].

5.2.1 Conventional Approach and Limitation

In [96], an algorithm is proposed to search such a segmentation $\bar{i}$ for the given sequence $\boldsymbol{x}$ that minimizes the cost $\beta_{\mathrm{D}}(\bar{i}, c)$ with a fixed tree penalty constant $c = 0.5$. However, $c = 0.5$ is proposed for estimating a context tree for infinite stationary sequence, and not for a non-stationary

sequence with segments where each segment is characterized by a distinct context tree. Moreover, our synthetic data results show that $c = 0.5$ sometimes proves to be higher and results in a lesser number of segments than the actual. Therefore, under our proposed criterion for segmentation, presented in Section 5.2.2, we consider $c \leq 0.5$.

5.2.2 Our Proposed Approach

Let $\overline{K}$ be a list of values of $c \leq 0.5$. Assuming that $c \leq 0.5$ is uniformly distributed, the average BIC cost or score of a segmentation $\bar{i}$, for the given data sequence, is given by

$$\beta_{\mathrm{A}}(\bar{i}) = \frac{1}{|\overline{K}|} \sum_{c \in \overline{K}} \beta_{\mathrm{D}}(\bar{i}, c). \tag{5.15}$$

To evaluate this average score, we fix the segmentation boundaries under a given segmentation i of the data sequence, vary the tree penalty constant c, and evaluate the score $\beta_{\mathrm{D}}(\bar{i}, c)$ for each c. Finally we take the average of the scores over all values of c. Note that for every fixed value of c, we estimate the context trees for all segments to get the segments' BIC scores for that c. Our proposed segmentation criterion is to find a segmentation $\bar{i}$ for the given sequence $\boldsymbol{x}$ that minimizes this average BIC cost, i.e.,

$$\overline{i^*} = \underset{\bar{i} \in \mathcal{H}_{\mathrm{S}}}{\arg\min}\, \beta_{\mathrm{A}}(\bar{i}) \tag{5.16}$$

Once the segmentation is finalized, we then estimate the context tree for each segment. Let $\overline{T^*}$ be the list of estimated context trees for all the segments in $\overline{i^*}$. The tree for the kth segment is estimated by minimizing the average BIC score of the segment, averaged over all c in $\overline{K}$, i.e.,

$$T_k^* = \underset{\overline{\tau} \in \mathcal{F}_D(\boldsymbol{x})}{\arg\min} \frac{1}{|\overline{K}|} \sum_{c \in \overline{K}} B_{\mathrm{S}}(x_{i_{k-1}^*+1}^{i_k^*}, \overline{\tau}, c). \tag{5.17}$$

Recall that D denotes the maximum possible depth of the tree. According to (5.7), for a given tree, $1/|\overline{K}| \sum_{c \in \overline{K}} B_{\mathrm{S}}(x_{i_{k-1}+1}^{i_k}, \overline{\tau}, c) = B_{\mathrm{S}}(x_{i_{k-1}+1}^{i_k}, \overline{\tau}, c_{\mathrm{av}})$, where $c_{\mathrm{av}} = \sum c/|\overline{K}|$. Thus

$$T_k^* = \underset{\overline{\tau} \in \mathcal{F}_D(\boldsymbol{x})}{\arg\min} B_{\mathrm{S}}(x_{i_{k-1}^*+1}^{i_k^*}, \overline{\tau}, c_{\mathrm{av}}). \tag{5.18}$$

Our goal in this study is to find the segmentation $\overline{i^*}$ and the trees $\overline{T^*}$, for a given sequence $\boldsymbol{x}$.

5.3 A Suboptimal Segmentation Algorithm

The search for the optimal segmentation $\overline{i^*}$ over the entire set $\mathcal{H}_{\mathrm{S}}$ is computationally very expensive. Therefore, instead of searching over $\mathcal{H}_{\mathrm{S}}$, for a given maximum possible number of segments $N_{\max}$, we search over a *reduced hypothesis set* of candidate segmentations, $\mathcal{H}_{\mathrm{SR}}(N_{\max})$, given by

$$\mathcal{H}_{\mathrm{SR}}(N_{\max}) = \left\{\overline{i_{\mathrm{B}}}(c)\middle| c \in \overline{K}\right\}, \tag{5.19}$$

$$\overline{i_{\mathrm{B}}}(c) = \underset{\bar{i}\in\hat{\mathcal{H}}_{\mathrm{S}}(N_{\max})}{\arg\min}\ C(c,\bar{i}). \tag{5.20}$$

Note that $\hat{\mathcal{H}}_{\mathrm{S}}(N_{\max})$ is a set of all possible segmentations of $\boldsymbol{x}$ with at the most $N_{\max}$ segments. Recall that $C(c,\bar{i}) = \beta_{\mathrm{D}}(\bar{i},c)$ is the BIC cost of a segmentation for a given c, given by (5.14). We call $\overline{i_{\mathrm{B}}}(c)$ as the ***BIC compliant*** segmentation for a given c, and $\mathcal{H}_{\mathrm{SR}}(N_{\max})$ as the set of BIC compliant segmentations.

For a large given $N_{\max}$, (5.20) is computationally very expensive to solve. Thus, we devise a *greedy binary segmentation* procedure to solve (5.20) in order to find a BIC compliant segmentation for a given c. This procedure is discussed in detail in Section 5.4. The overall segmentation procedure is given in Figure 5.2. Note that the Greedy_Binary_Segmentation procedure is explained in Section 5.4. Also, $l_{\min}$ represents the minimum segment length, which is a given parameter, explained in Section 5.4.2.

5.4 Greedy Binary Segmentation

For a given tree penalty constant c, we find the BIC compliant segmentation of the given sequence $\boldsymbol{x}$ according to the procedure shown in Figure 5.3. We begin with an initial segmentation $\bar{i}_0$ of the given sequence $\boldsymbol{x}$, representing only one segment. Then we keep splitting $\boldsymbol{x}$ iteratively using the Greedy_Binary_Split procedure, described in Section 5.4.1, until we reach the given maximum number of segments $N_{\max}$. We keep track of all these segmentations, i.e., $\bar{i}_k$ for all $k < N_{\max}$, and compute the cost corresponding to each one of them. Finally we select the segmentation which has the minimum cost. Recall that the cost for a given segmentation is given by (5.14).

Segment(Δ_{c}, $l_{\min}$, $N_{\max}$, D):

- $\overline{K} = [\Delta_{\mathrm{c}}, 2\Delta_{\mathrm{c}}, \ldots, 0.5]$
- For $c \in \overline{K}$:

  $\overline{i_{\mathrm{B}}}(c)$ = Greedy_Binary_Segmentation(c, $N_{\max}$, $l_{\min}$, D)
- Build the set $\mathcal{H}_{\mathrm{SR}}(N_{\max}) = \left\{\overline{i_{\mathrm{B}}}(c) \middle| c \in \overline{K}\right\}$
- Find $\overline{i^*} = \arg\min_{\bar{i} \in \mathcal{H}_{\mathrm{SR}}(N_{\max})} \beta_{\mathrm{A}}(\bar{i})$
- For every kth segment in $\overline{i^*}$:

  $\boldsymbol{x}_{\mathrm{seg}} = x_{i^*_{k-1}+1}^{i^*_k}$

  T^*_k, P^*_k = Estimate_Context_Tree(c_{av}, $\boldsymbol{x}_{\mathrm{seg}}$, D)
- Return $\overline{i^*}$, $\overline{T^*}$, $\overline{P^*}$

FIGURE 5.2: The overall segmentation procedure.

5.4.1 Greedy Binary Split Procedure

This procedure bisects one of the segments of the input segmentation $\bar{i}$ and returns the updated segmentation. The details of the procedure are given in Figure 5.4. This procedure does not bisect just any segment, rather it selects an appropriate segment to split. To this end, the procedure finds all possible binary splits, $\overline{v}_k$ for all $k < |\bar{i}|$, for the given segmentation $\bar{i}$. Each kth binary split, represented as a list of updated segment end indices $\overline{v}_k$, is obtained after the bisection of the kth segment of the given sequence $\boldsymbol{x}$ under segmentation $\bar{i}$. Finally the split with the minimum cost is selected and returned.

Note that a segment is greedily split using the BIC_Bisection procedure, described in Section 5.4.2. Here, 'greedy' means that a given segment is bisected under the BIC criterion which results in a bisection with a BIC score that is optimum (minimum) locally for the given segment, independent of the rest of the segments. Moreover, as explained in detail in Section 5.4.2,

Greedy_Binary_Segmentation(c, $N_{\max}$, $l_{\min}$, D)**:**

- $\bar{i}_0 = \left[l(\boldsymbol{x}) - 1\right]$
- $\mathcal{N} = \left\{0, 1, 2, \ldots, N_{\max} - 1\right\}$
- For $k \in \mathcal{N} \setminus \{0\}$:

  $\bar{i}_k$ = Greedy_Binary_Split(c, $\bar{i}_{k-1}$, $l_{\min}$, D)
- $\mathcal{C} = \left\{C(c, \bar{i}_k) \middle| k \in \mathcal{N}\right\}$
- $\overline{i_{\mathrm{B}}} = \bar{i}$, s.t. $C(c, \bar{i}) = \min \mathcal{C}$
- Return $\overline{i_{\mathrm{B}}}$

FIGURE 5.3: The greedy binary segmentation procedure.

BIC_Bisection procedure limits the minimum segment length to be $l_{\min}$ number of symbols, therefore the Greedy_Binary_Split procedure only bisects a segment if its length is at least $2l_{\min}$.

5.4.2 BIC Bisection Procedure

For a given c, this procedure bisects a given segment of $\boldsymbol{x}$, which is specified by the segment start and end indices, i.e., i_{start} and i_{end} respectively, as shown in Figure 5.5. The procedure returns i_{bisect}, i.e., the segment end index of the first bisection half that resulted after bisection. Note that the given segment is bisected such that the sum of the BIC scores of both the resulting bisection halves is the minimum of those of all possible bisections. Moreover, the procedure limits the minimum segment length to be $l_{\min}$ number of symbols. It is required because we estimate the context tree for each segment for evaluating its BIC score, as indicated by (5.13). Thus, we need to set a sufficient limit on the minimum number of data points (symbols) in a segment, in order to have a reliable estimate of the context tree and the corresponding BIC score.

5.5 Results based on Synthetic Data

To validate the algorithm presented in Section 5.3, we conduct experiments based on synthetic data. In these experiments, we generate sequences with known segmentation and context tree models, and then run our segmentation algorithm to detect the segmentation boundaries. In Sec-

Greedy_Binary_Split(c, $\bar{i}$, $l_{\min}$, D):

- $\mathcal{N}$ = $\{0, 1, 2, \ldots, |\bar{i}| - 1\}$
- For $k \in \mathcal{N}$:

  $\overline{v}_k$ = Copy($\bar{i}$)

  $l_k = i_k - i_{k-1}$

  If $l_k \geq (2 \times l_{\min})$:

    b = BIC_Bisection(c, $i_{k-1} + 1$, i_k, $l_{\min}$, D)

    Insert b in $\overline{v}_k$ at index k
- $\mathcal{C}$ = $\left\{C(c, \overline{v}_k) \middle| k \in \mathcal{N}\right\}$
- $\overline{v}_{\text{opt}}$ = $\overline{v}$, s.t. $C(c, \overline{v}) = \min \mathcal{C}$
- Return $\overline{v}_{\text{opt}}$

FIGURE 5.4: The greedy binary split procedure.

tions 5.5.1 and 5.5.2, we present the details of these experiments along with a comparison between the actual and the detected segmentation boundaries. The tree models given in Table 5.1 are used to generate the sequence data in the experiments, as detailed in Sections 5.5.1 and 5.5.2. The graphical representation of one of these original tree models is also shown in Figure 5.6. Note that the complete list of the alphabets in our experiments is $[0, 1, 2, 3, 4]$.

TABLE 5.1: Context Tree Models for Generating Data

Context lists	Transition probability lists
$T_{\text{A}} = [[0], [1], [3], [4]]$	$P_{\text{A}} = [[0, 0.310, 0, 0.241, 0.448], [1, 0, 0, 0, 0], [1, 0, 0, 0, 0], [1, 0, 0, 0, 0]]$
$T_{\text{B}} = [[0], [2], [4], [0, 1]]$	$P_{\text{B}} = [0, 0.248, 0.063, 0, 0.688], [1, 0, 0, 0, 0], [1, 0, 0, 0, 0], [1, 0, 0, 0, 0]]$
$T_{\text{C}} = [[0], [3], [4]]$	$P_{\text{C}} = [[0, 0, 0, 0.5, 0.5], [1, 0, 0, 0, 0], [1, 0, 0, 0, 0]]$
$T_{\text{D}} = [[0], [4]]$	$P_{\text{D}} = [[0, 0, 0, 0, 1], [1, 0, 0, 0, 0]]$

5.5.1 Same Length Segments

We generate a 10 segment sequence from the alphabets $[0, 1, 2, 3, 4]$, shown as the first column in Figure 5.7, where each segment is of length 200 symbols. The 1st, 5th and 9th segments are

BIC_Bisection(c**,** i_{start}**,** i_{end}**,** $l_{\min}$**,** D**):**

- $j_{\text{i}} = i_{\text{start}} + l_{\min} - 1$
- $j_{\text{f}} = i_{\text{end}} - l_{\min}$
- $J = \{j_{\text{i}}, j_{\text{i}} + 1, \ldots, j_{\text{f}}\}$
- For $j \in J$:

$$\beta_{\text{sum}}(j) = \beta_{\text{S}}\big(x_{i_{\text{start}}}^{j}, c, D\big) + \beta_{\text{S}}\big(x_{j+1}^{i_{\text{end}}}, c, D\big)$$

- $i_{\text{bisect}} = \arg\min_{j \in J} \beta_{\text{sum}}(j)$
- Return i_{bisect}

FIGURE 5.5: The BIC bisection procedure.

generated using context tree T_{A}. The 2nd, 6th and 10th segments are generated using context tree T_{B}. The 3rd and 10th segments are generated using context tree T_{C}. The 4th and 8th segments are generated using context tree T_{D}. In Figure 5.7, we also present the set $\mathcal{H}_{\text{SR}}$ of the BIC compliant candidate segmentations for the generated sequence. Recall that this set $\mathcal{H}_{\text{SR}}$ is described in Section 5.3. Each of this candidate is generated for every $c \in \overline{K}$, for $\Delta_{\text{c}} = 0.05$, using the Greedy_Binary_Segmentation procedure presented in Section 5.4, with $N_{\max} = 200$, $l_{\min} = 5$ and $D = 5$.

Figure 5.8 shows the plots of the average BIC score β_{A}, and the distance of each candidate segmentation from the original. Recall that β_{A} for a given segmentation is given by (5.15), and it is the score that we is used in our proposed algorithm to select the final segmentation out of the BIC compliant candidates, as described in the Segment procedure in Section 5.3. Note that we use the same metric to find the distance between two segmentation models as that proposed in [96]. This distance measure is given by

$$S(\bar{i}, \overline{v}) = \max\{\tilde{S}(\bar{i}, \overline{v}), \tilde{S}(\overline{v}, \bar{i})\}. \tag{5.21}$$

$$\tilde{S}(\bar{i}, \overline{v}) = \frac{1}{|\bar{i}|} \sum_{k=-1}^{|\bar{i}|-1} \min_{v_l \in [-1, \overline{v}]} \left\{ \frac{|i_k - v_l|}{l(\boldsymbol{x})} \right\} \tag{5.22}$$

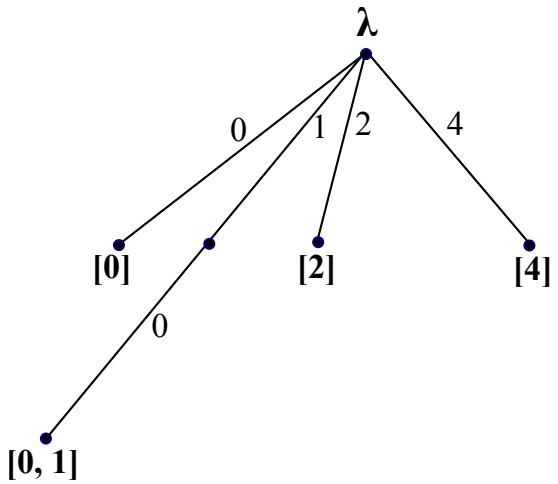


FIGURE 5.6: Context tree T_{B} from Table 5.1.

The measure $\tilde{S}(\bar{i}, \overline{v})$ is the distance of each segment boundary or segment end index in segmentation $\bar{i}$ to the closest segment boundary in segmentation $\overline{v}$ on average. This distance is also normalized to the total length of the data sequence $\boldsymbol{x}$, and thus ranges between 0 and 1. Figure 5.8 clearly shows that the candidate segmentation that minimizes β_{A} is very close to the original segmentation. For this experiment the candidate segmentations that minimizes β_{A} are the ones corresponding to $c = 0.3$ and $c = 0.35$. Any one of these can be selected as the final segmentation for the generated sequence. The estimated context trees for the first five segments of the final selected segmentation are compared to the original trees in Table 5.2.

TABLE 5.2: Context Tree Models for Experiment Results of Figure 5.7

Original context lists	Estimated context lists (T_k^*) for first five segments
$T_{\mathrm{A}} = [[0], [1], [3], [4]]$	$T_0^* = [[0], [1], [3], [4]]$
$T_{\mathrm{B}} = [[0], [2], [4], [0, 1]]$	$T_1^* = [[1], [2], [4], [1, 0], [4, 0], [0, 2, 0]]$
$T_{\mathrm{B}} = [[0], [2], [4], [0, 1]]$	$T_2^* = [[0], [1], [4]]$
$T_{\mathrm{C}} = [[0], [3], [4]]$	$T_3^* = [[0], [3], [4]]$
$T_{\mathrm{D}} = [[0], [4]]$	$T_4^* = [[0], [4]]$

5.5.2 Different Length Segments

We generate a 3 segment sequence from the alphabets $[0, 1, 2, 3, 4]$, shown as the first column in Figure 5.9, with different segment lengths. The 1st and 3rd segments are generated using the con-

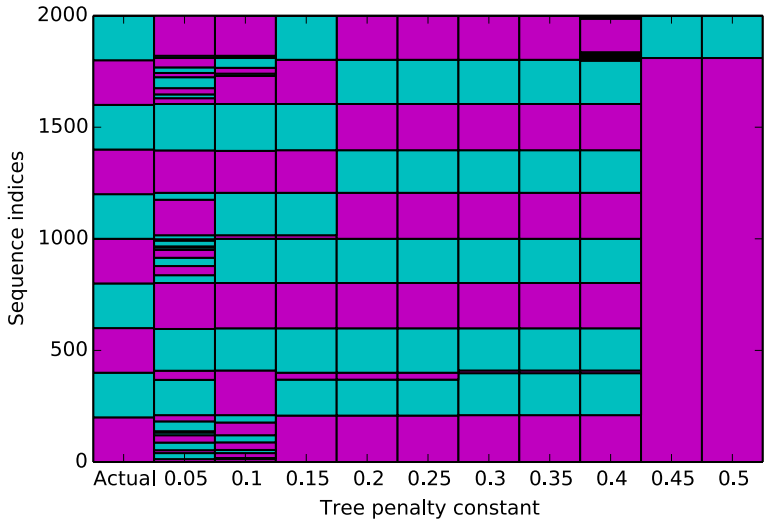


FIGURE 5.7: BIC compliant candidate segmentations. The finally selected segmentation with the minimum score β_{A} is for $c = 0.3$ and $c = 0.35$.

text tree T_{A}. The 2nd segment is generated using the context tree T_{B}. Similar to Section 5.5.1, we present the set $\mathcal{H}_{\mathrm{SR}}$ of the candidate segmentations in Figure 5.9. Note that we use the same values for the input parameters of the segmentation algorithm as those in Section 5.5.1. Figure 5.10 clearly shows that the candidate segmentation that minimizes β_{A} is very close to the original segmentation, similar to the experimental results presented in Section 5.5.1. For this experiment the candidate segmentation that minimizes β_{A} is the one corresponding to $c = 0.15$.

5.6 Results based on Public Safety Communications Data

In this study, we use the call records meta-data of the public safety communication system (PSCS) of the state of Louisiana. PSCSs are implemented as trunked mobile radio systems. In these systems, the users are divided into fleets or groups normally called *talk groups*. Each user can only communicate with another user of the same talk group. When a user needs to talk, it presses the push-to-talk (PTT) button of its radio, in order to send a call request at the site's *base-station*. The base-station assigns communication resources to the caller and the rest of the users of the talk group, in order to broadcast the voice call throughout the group.

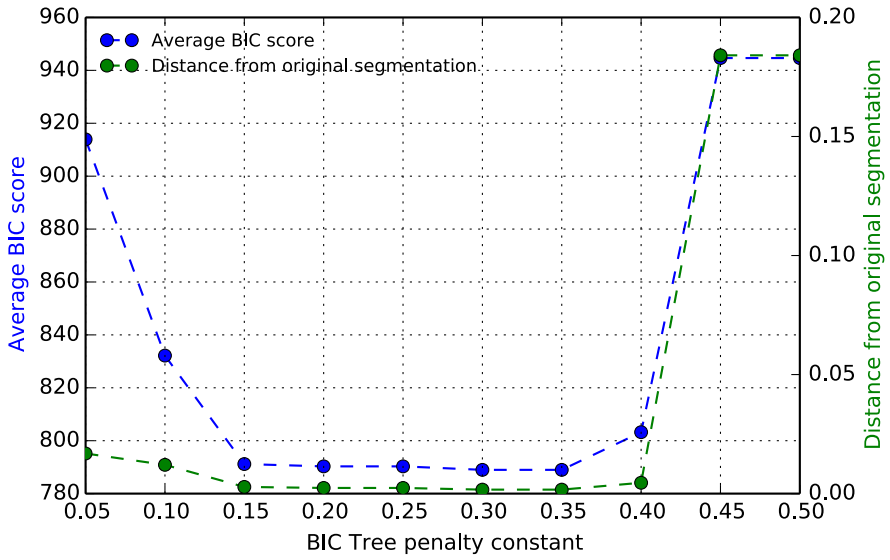


FIGURE 5.8: Average BIC score β_{A}, and the distance from the original segmentation, for the candidates shown in Figure 5.7.

5.6.1 Talk Group Activity and Communication Behavior

We quantify a talk group's communication *activity* as a discrete sequence of symbols $[0, 1, 2, 3, 4]$, which we extract from the call records data. Symbol 1 represents that the caller radio is the busiest caller in the talk group. Similarly, symbol 2 represents the 2nd busiest, and symbol 3 shows the 3rd busiest caller radio. Symbol 4 represents that the caller radio is one of the rest of the radios. Symbol 0 is used to represent an idle duration, where there is no communication activity by the talk group. Note that the call durations in PSCSs are very small and are of the order of seconds. However, an idle duration can be very large in PSCSs, as compared to the call durations, and can be of the order of hours. Therefore, we also incorporate the idle durations while extracting the activity sequence of a talk group. To this end, we up-sample an idle symbol by a factor of $\lceil$idle duration / average idle duration$\rceil$, if idle duration $>$ average idle duration. We then apply our segmentation algorithm on these discrete activity sequences.

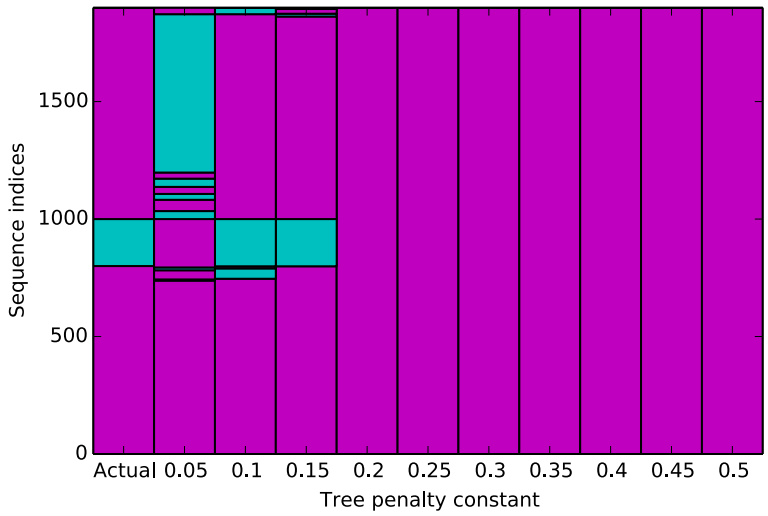


FIGURE 5.9: BIC compliant candidate segmentations. The finally selected segmentation with the minimum score β_{A} is for $c = 0.15$.

5.6.2 Fire Incidents Data

Note that we do not have any 'ground truth' to validate our segmentation results for the real public safety communications data. Therefore, we use the fire incidents timeline to validate the segmentation of a talk group's activity sequence which is suggested by our proposed segmentation algorithm, as explained in Section 5.6.4. We use the fire incidents data since it is available online as an open source data set [97]. The first column in Figures 5.11 and 5.12 represents the occurrences of fire incidents in Baton Rouge, on January 26, 2014, from 00:00:00 to 23:59:59. A cyan colored window in the first column shows the duration of a fire incident. We generate this incidents timeline using the open source data available online [97].

5.6.3 Group Selection Criteria

We have used two criteria to select the appropriate talk groups to analyze their call records data using our segmentation algorithm. Under the first criterion, we select the fire department talk groups out of the busiest talk groups based on the whole day activity. CBR-VF-DISP-1 is one such group whose results are presented in Figure 5.11. Under the second criterion, we select such talk groups that are the busiest groups based on the activity within a fire incident time window.

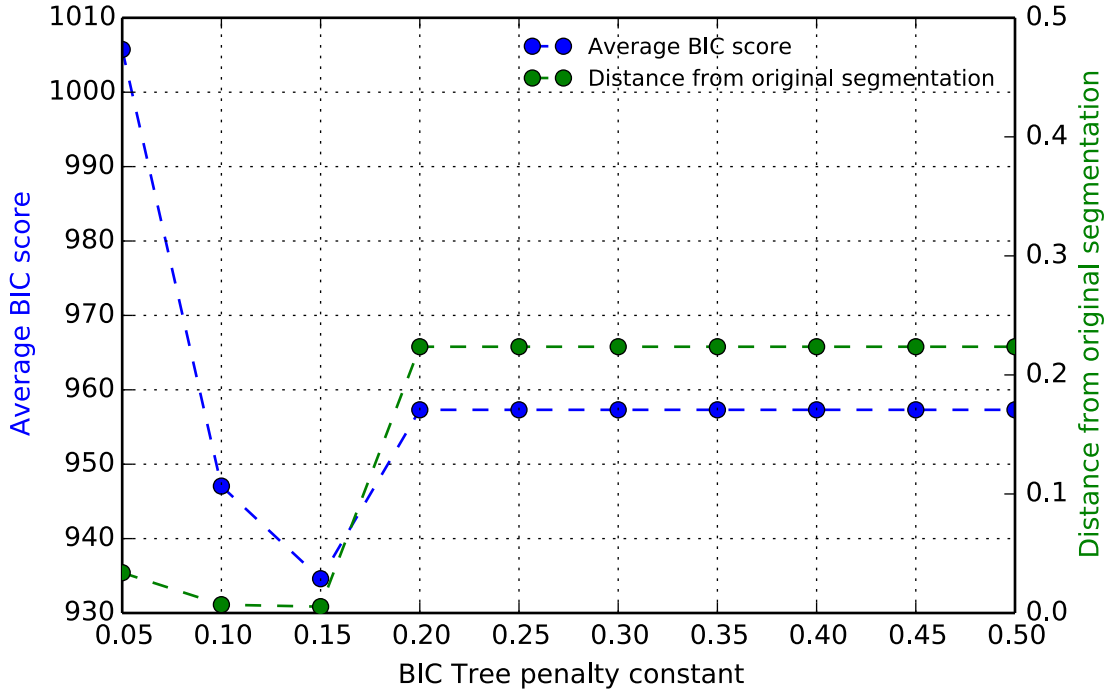


FIGURE 5.10: Average BIC score β_{A}, and the distance from the original segmentation, for the candidates shown in Figure 5.9.

CBR-D2-DISP is one such group whose results are presented in Figure 5.12. Note that this is one of the busiest group during the second fire incident.

5.6.4 Segmentation Results

In Figures 5.11 and 5.12, the second column represents the segmentation results for the activity sequences of two talk groups, namely, CBR-VF-DISP-1 and CBR-D2-DISP, for January 26, 2014, from 00:00:00 to 23:59:59. These groups are selected under the criteria explained in Section 5.6.3. For segmentation, we use the Segment procedure described in Section 5.3, with parameter values as follows, $\Delta_{\mathrm{c}} = 0.05$, $N_{\max} = 200$, $l_{\min} = 30$ and $D = 5$. A change in color in the second column represents a change in the *communication behavior* of the talk group. Note that the communication behavior of the talk group during an activity segment is quantified as a VLMC of the activity symbols, represented as a context tree which is estimated for that segment. In Figures 5.11

and 5.12, we present the segmentation in terms of time indices, in contrast to Section 5.5 where we present the results in terms of sequence symbol indices.

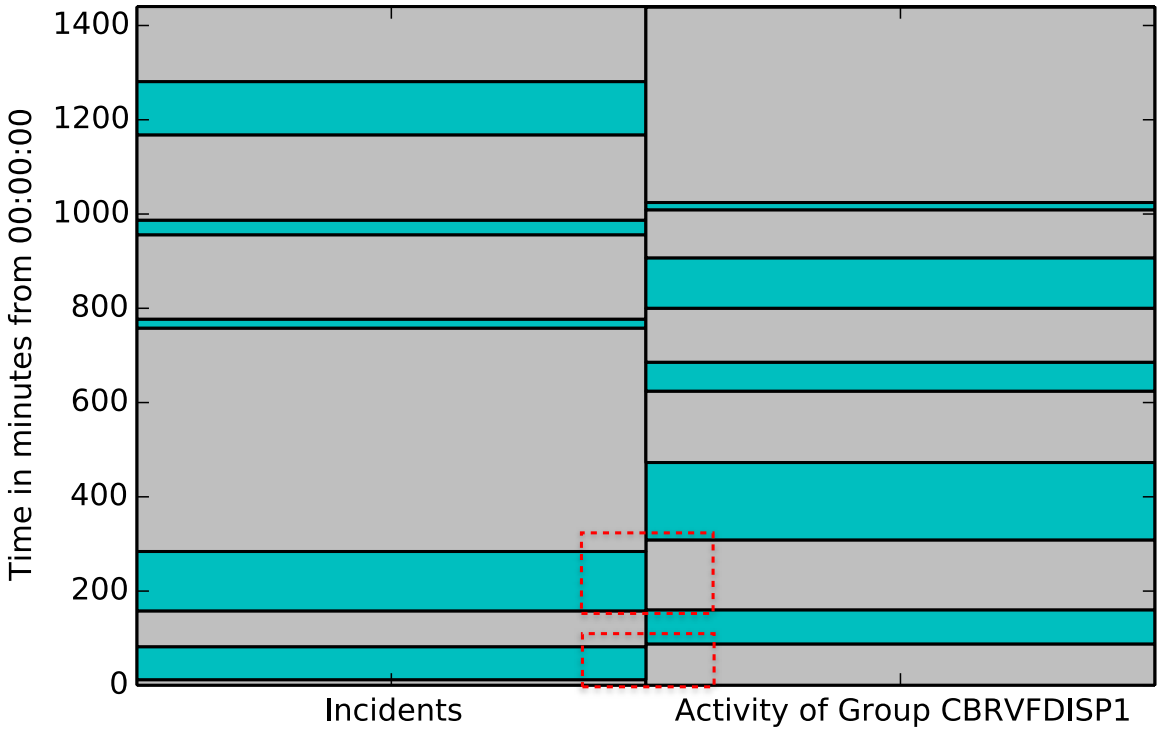


FIGURE 5.11: Activity segmentation of talk group CBR-VF-DISP-1.

The beginning and ending of an incident window in the first column suggests a change in the communication behavior of the talk group. A comparison of both the columns in Figures 5.11 and 5.12 shows that a change in communication behavior of the talk group, as suggested by the beginning and ending of an incident, is captured by our segmentation algorithm. In Figure 5.11 activity segments are detected corresponding to the 1st and 2nd incidents. In Figure 5.12 an activity segment is detected by our segmentation algorithm corresponding to the 2nd incident in the first column. For clarity, we have used dotted red boxes around the activity segments in the second column and their corresponding incidents windows in the first column. Also, the graphical representations of two of the estimated context trees for the group CBR-VF-DISP-1 are also shown in Figure 5.13, for the segments at indices 6 and 7.

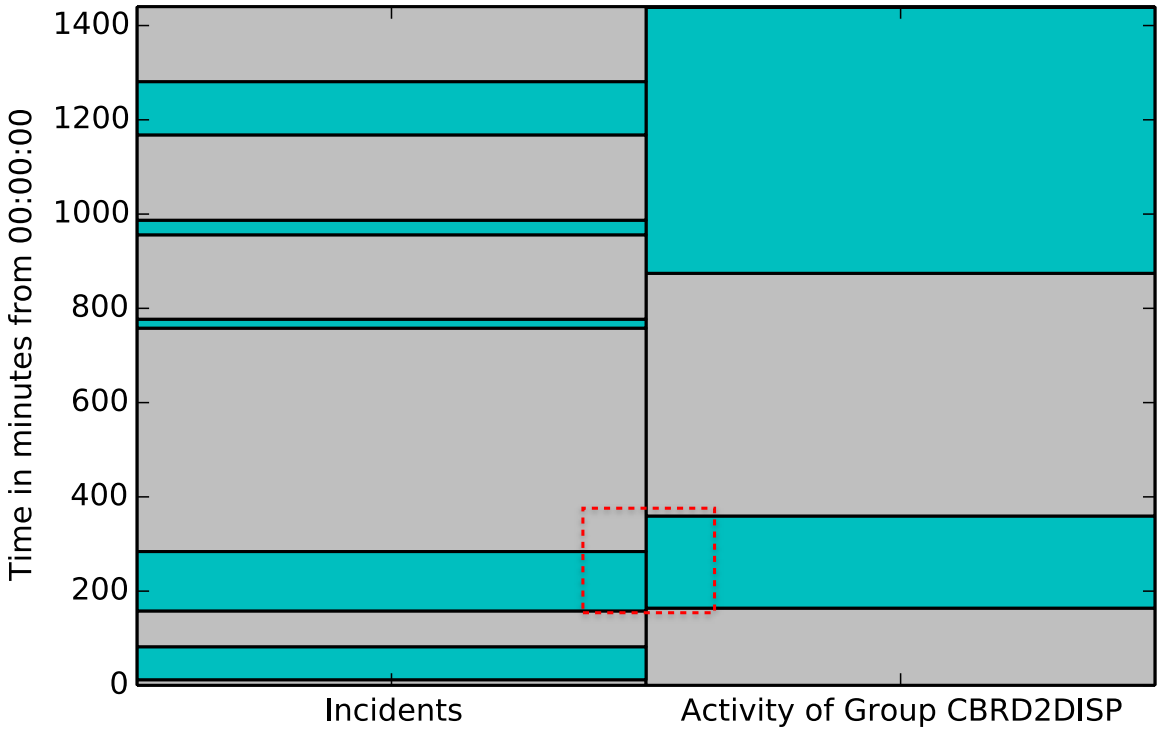


FIGURE 5.12: Activity segmentation of talk group CBR-D2-DISP.

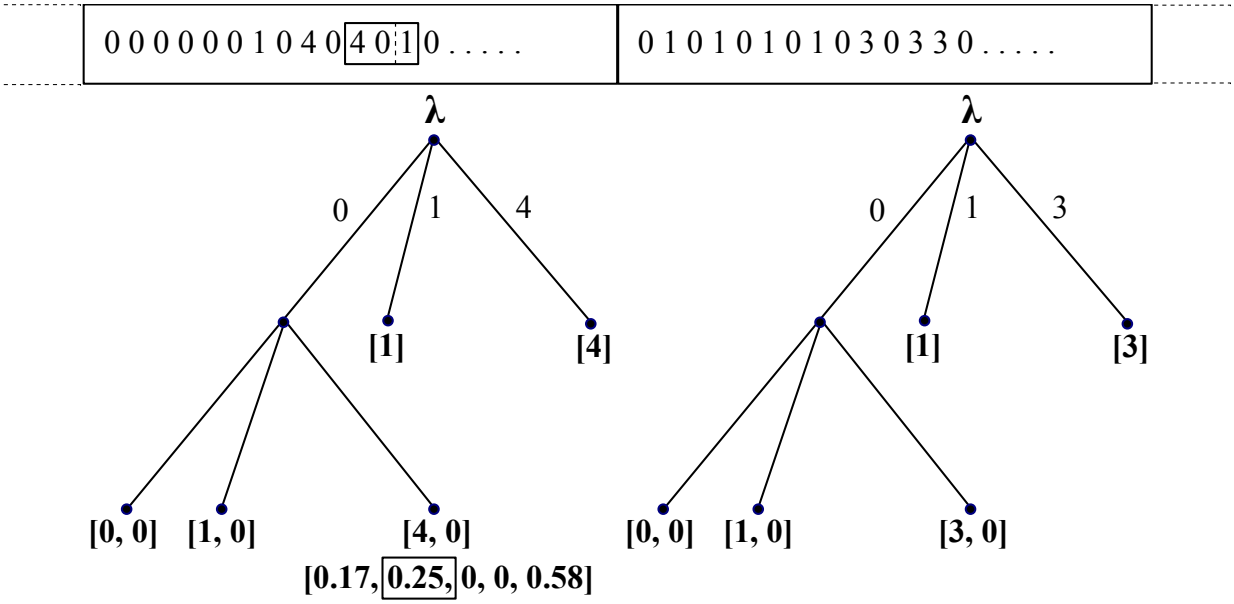


FIGURE 5.13: Context trees T_6^* and T_7^* for the group CBR-VF-DISP-1.

Chapter 6
Summary and Future Works

The goal of our study is to propose techniques for an efficient allocation of spectral resources in finite source systems. These techniques can help alleviate the overloading of public safety communication systems during emergency situations. Our study consists of four projects that are discussed in detail in Chapters 2, 3, 4 and 5. In this chapter, we present a brief summary of our contributions and discuss interesting future research ideas for each of these projects.

6.1 Spectrum Allocation via Cognitive Radio based System

In Chapter 2, we study how an efficient spectrum allocation can be achieved by adopting a *cognitive radio* based design approach. In a cognitive radio based system, there are two types of users, namely, primary users (PUs) and secondary users (SUs). PU is a high priority user that can use the allocated frequency channel at any time when needed. On the other hand, SU is a low priority user that can use the channel allocated for PU, when PU is idle and the channel is available for SU. One of the challenges faced while designing such systems is to improve the capability of the SU transmitter (SU-TX) to detect the availability of the channel. This can be achieved by appropriately selecting the detection parameters of SU-TX's signal detector that result in an optimal *sensing-throughput tradeoff*, as explained in Chapter 2. In that chapter, we provide the design guidelines for the cognitive radio systems, based on the transmission-power levels of the users. Note that the design goal in this project is to achieve the maximum SU performance by keeping the PU performance-degradation within a given tolerable range. Some interesting future research ideas are as follows.

- The system model discussed in Chapter 2 does not incorporate *fading*. However, we briefly discuss the effect of fading on the degrees of freedom and interference trade-off for cognitive radio systems in Section 2.6.1. A more detailed study that includes the incorporation of

fading in the system model, and its effect on the PU and SU performances, is left as a part of the future work, that will provide further useful insights.

- Also, in Chapter 2, we only consider a single point-to-point communication link for both PU and SU in our system model. Even this simple model poses analytical challenges and reveals interesting results, as elaborated in this study. However, this basic model can be used as a building block for a more sophisticated model of cognitive radio networks, which may consist of large number of nodes and communication links among them. Such complicated model considerations are also left as a part of the future work.

6.2 Communication Channel Allocation by Call Admission and Preemption Control

In Chapter 3, we study the communication channel allocation as the call admission and preemption control, for finite source systems. To this end, the system is modeled as a prioritized queueing system, wherein the users are assigned priority levels, and the system operation is modeled as a Markov decision process (MDP). Based on these priorities, optimal decisions are determined for admitting a new call or preempting an already busy call. These policy decisions depend on the priority level of the call and the state of the system. In this study, we demonstrate that the optimal policies are *threshold based* policies. Note that the threshold based policies are easy to design and can be implemented in terms of decision thresholds. These decision thresholds are used to make optimal admission and preemption decisions during the system operation. Some interesting future research ideas are as follows.

- In Chapter 3, we consider time-invariant call arrival and service rates. An interesting scenario to analyze as a part of the future work is to consider time varying call arrival and service rates.

- Also, in Chapter 3, we demonstrate using numerical results that the optimal low priority call preemption policy is a threshold based policy. However, the analytical proof of this

observation is left as a part of the future work. We can adopt a similar approach to prove this result as that adopted for the admission policy.

6.3 Optimal Joint Allocation of Control and Communication Channels

In a control channel based wireless access system, e.g., a public safety communication system (PSCS), there are two types of frequency channels. One of the types is called the control channel that is dedicated for access control, whereas the rest of the channels are used for communications. Therefore, we can divide the system into two segments or layers, namely, the access layer and the communication layer. In order to implement a bandwidth efficient system we need an optimal allocation of control and communication channels. As explained in detail in Chapter 4, the performance of one layer significantly affects that of the other one, in case of the finite source wireless systems like PSCSs. Thus, for such systems, the control and communication layers are inseparable. Therefore, in Chapter 4, we propose a novel statistical model for wireless access systems that jointly models the control and communication layers, and helps evaluate the optimal number of control and communication channels. We also propose the concept of a *channel allocation map* that helps visualize the optimal channel allocation for all possible values of the call traffic parameters. Note that the optimal channel allocation also requires the knowledge of the actual call-arrival traffic load. However, this load is invisible to a practical system, because in practice, a system does not keep records of the calls that are blocked due to collision at the access layer. Therefore, in Chapter 4, we also demonstrate the capability of our proposed model in estimating the invisible actual traffic load. Finally, we provide guidelines for developing an algorithm for the traffic aware allocation of channels, based on our proposed model. Some interesting future research ideas are as follows.

- In Chapter 4, we assume that the physical layer of the system is simple and perfect, and does not cause any deterioration in communications. Incorporation of a more sophisticated physical layer in our system model and analysis, by considering fading and path-loss models, is left as a part of the future work.

- Also, in Chapter 4, we assume that the control and the communication channels provide the same quality of service. Consideration of different qualities of control and communication channels is also left as a part of the future work.

6.4 Segmentation of Talk Group's Call Activity

The spectrum allocation techniques, e.g., the ones discussed in Chapters 2, 3 and 4, requires the knowledge of the call traffic parameters and the priority levels of the users in the system. For practical systems, these required pieces of information are extracted from the call records meta-data. A key fact that should be considered while analyzing the call records is that the call arrival traffic and the users priority levels change with a change in events on the ground. This is so because a change in events on the ground affects the communication behavior of the talk groups in the system, which affects the call arrival traffic and the priority levels of the users. Thus, the first and the foremost step in analyzing the call records data for a given talk group, for extracting the call traffic information, is to segment the data into time intervals of homogeneous or stationary communication behavior of the group. A mechanism for such a segmentation of the data of a practical PSCS is discussed in Chapter 5.

In Chapter 5, we develop a way to quantify a talk group's activity as a discrete sequence of symbols. Here a symbol, corresponding to a call, represents the calling radio-unit in the talk group, as explained in Chapter 5. Note that we extract this talk group activity from the available data of call records. In practice, a talk group's communication behavior remains consistent during a certain event happening on the ground. However, with a change in event, the talk group's behavior also changes. Thus, a talk group's activity, extracted as a discrete sequence, from the whole day data of call records, is not stationary. However, it may consist of many stationary segments depending on different events that occurred on the ground. Therefore, in Chapter 5, we also propose an algorithm to segment a given whole day activity of a talk group into stationary segments, each corresponding to a distinct event, and also quantify the behavior of the talk group within each segment or event. Note that we quantify the behavior of a talk group in the form of a *context*

tree which represents a *variable memory Markov process* of discrete symbols, as explained in Chapter 5. Some interesting future research ideas for this project are as follows.

- As a part of the future work, we are interested in developing a metric to quantify the difference between the context trees estimated for the adjacent segments. In this way, we can refine our segmentation results by merging the similar segments together, if their context trees are not very different from each other.

- In Chapter 5, we have only considered the communications within the same given talk group. However, as a part of the future work, we are also interested in characterization and analysis of the interaction among different talk groups. The main motivation for this prospective future work is to develop mechanisms for determining the talk groups' priority levels based on their interactions with one another. Note that these priority levels are required for the resource allocation mechanisms, e.g., the ones discussed in Chapters 2 and 3.

- Also, the segmentation algorithm developed in Chapter 5 is of off-line nature. To learn the current ongoing contexts, an online algorithm with the features of sequential and context identification should be investigated. The off-line algorithms, such as the one proposed in Chapter 5, will allow us to build incidents based lists of high priority groups, also called the site affiliation templates, in terms of the featured context trees. As a contrast, the online algorithms will allow us to identify the variation of contexts, as well as the resulting priority orders of different talk-groups, while these groups are working in the fields. Such knowledge of the priority orders of the groups will be of tremendous help in allocating network resources in a context-aware manner.

Bibliography

[1] D. Čabrić, S. M. Mishra, D. Willkomm, R. Brodersen, and A. Wolisz, "A cognitive radio approach for usage of virtual unlicensed spectrum," in *Proc. of 14th IST Mobile Wireless Communications Summit*, 2005, pp. 1–4.

[2] J. Mitola, "Cognitive radio for flexible mobile multimedia communications," in *IEEE Int'l. Wksp. Mobile Multimedia Communications*, San Diego, CA , USA, November 1999.

[3] E. Hossain, D. Niyato, and D. I. Kim, "Evolution and future trends of research in cognitive radio: a contemporary survey," *Wireless Communications and Mobile Computing*, 2013.

[4] R. Ferrus, O. Sallent, G. Baldini, and L. Goratti, "Public safety communications: Enhancement through cognitive radio and spectrum sharing principles," *IEEE Vehicular Technology Magazine*, vol. 7, no. 2, pp. 54–61, June 2012.

[5] C. Chen, R. Hudson, and K. Yao, "Modeling and theoretical performance analysis for dynamic spatially distributed energy-based spectrum sensing in cognitive radios over shadowed fading channels in public safety networks," in *IEEE Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, June 2012, pp. 1–6.

[6] A. Gorcin and H. Arslan, "Public safety and emergency case communications: Opportunities from the aspect of cognitive radio," in *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, Oct 2008, pp. 1–10.

[7] N. Jesuale and B. Eydt, "A policy proposal to enable cognitive radio for public safety and industry in the land mobile radio bands," in *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, April 2007, pp. 66–77.

[8] A.-A. Ali and S. Wei, "Degrees-of-freedom and interference trade-off for a mix-Gaussian cognitive-radio channel," in *IEEE Annual Conference on Information Sciences and Systems (CISS)*, 2013.

[9] ——, "Analysis of degrees of freedom under mixture Gaussian model in cognitive radio systems," *Physical Communication*, 1st revision submitted in March 2015.

[10] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, 1998.

[11] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Dispersion of the Gilbert-Elliott channel," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 1829–1848, 2011.

[12] *Communication without channel state information at the receiver (CSIR)*, ([online] Available at: http://ita.ucsd.edu/wiki/index.php?title=Communication_without_Channel_State_Information_at_the_Receiver_(CSIR), accessed on March 2015).

[13] Y.-C. Liang, Y. Zeng, E. C. Y. Peh, and A. T. Hoang, "Sensing-throughput tradeoff for cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 4, pp. 1326–1337, 2008.

[14] H. Pradhan, S. Kalamkar, and A. Banerjee, "Sensing-throughput tradeoff in cognitive radio with random arrivals and departures of multiple primary users," *IEEE Communications Letters (2015)*.

[15] E. Bedeer, O. Dobre, M. Ahmed, and K. Baddour, "Rate-interference tradeoff in OFDM-based cognitive radio systems," *IEEE Transactions on Vehicular Technology (2014)*.

[16] A. Kortun, T. Ratnarajah, M. Sellathurai, Y.-C. Liang, and Y. Zeng, "On the eigenvalue-based spectrum sensing and secondary user throughput," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 3, pp. 1480–1486, 2014.

[17] S. Zhang, H. Zhao, S. Wang, and J. Wei, "A cross-layer rethink on the sensing-throughput tradeoff for cognitive radio networks," *IEEE Communications Letters (2014)*.

[18] S. K. Sharma, S. Chatzinotas, and B. Ottersten, "A hybrid cognitive transceiver architecture: Sensing-throughput tradeoff," in *International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)*. IEEE, 2014, pp. 143–149.

[19] M.-F. Hsu, T.-Y. Wang, and C.-T. Yu, "A unified spectrum sensing and throughput analysis model in cognitive radio networks," *Wireless Communications and Mobile Computing (2013)*, 2013.

[20] S. Zarrin and T. J. Lim, "Throughput-sensing tradeoff of cognitive radio networks based on quickest sensing," in *IEEE ICC*, Kyoto, Japan, June 2011.

[21] S. Stotas and A. Nallanathan, "Overcoming the sensing-throughput tradeoff in cognitive radio networks," in *IEEE ICC*, Cape Town, South Africa, May 2010.

[22] L. Tang, Y. Chen, E. L. Hines, and M.-S. Alouini, "Effect of primary user traffic on sensing-throughput tradeoff for cognitive radios," *IEEE Transactions on Wireless Communications*, vol. 10, no. 4, pp. 1063–1068, 2011.

[23] Y.-J. Choi, Y. Xing, and S. Rangarajan, "Overhead-throughput tradeoff in cooperative cognitive radio networks," in *IEEE WCNC*, Budapest, Hungary, April 2009.

[24] S. Huang, X. Liu, and Z. Ding, "Optimal-sensing transmission structure for dynamic spectrum access," in *IEEE INFOCOM*, Rio de Janeiro, Brazil, April 2009.

[25] Y. Pei, Y.-C. Liang, K. C. Teh, and K. H. Li, "Sensing-throughput tradeoff for cognitive radio networks: A multiple-channel scenario," in *IEEE PIMRC*, Tokyo, Japan, September 2009.

[26] E. C. Y. Peh, Y.-C. Liang, Y. L. Guan, and Y. Zeng, "Optimization of cooperative sensing in cognitive radio networks: A sensing-throughput tradeoff view," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 9, pp. 5294–5299, 2009.

[27] S. Wei, V. Chakravarthy, Z. Wu, and R. Kannan, "Sensing and transmission in probabilistically interference limited cognitive radio systems," in *IEEE GlobeCom 2011*, Houston, TX, USA, December 2011.

[28] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proceedings of the IEEE*, vol. 55, no. 4, pp. 523–531, 1967.

[29] F. F. Digham, M.-S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," *IEEE Transactions on Communications*, vol. 55, no. 1, pp. 21–24, 2007.

[30] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*. pg. 892, 899, Elsevier, 2007.

[31] A. Lapidoth, "Nearest neighbor decoding for additive non-gaussian noise channels," *IEEE Transactions on Information Theory*, vol. 42, no. 5, pp. 1520–1529, 1996.

[32] M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck, "On entropy approximation for gaussian mixture random vectors," in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, MFI 2008*, Seoul, August 2008, pp. 181–188.

[33] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley and Sons, 2006.

[34] H. L. V. Trees, *Detection, Estimation, and Modulation Theory - Vol 1*. John Wiley and Sons, 2001.

[35] C. Cordeiro, K. Challapali, D. Birru, and N. Sai Shankar, "IEEE 802.22: the first worldwide wireless standard based on cognitive radios," in *First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2005, pp. 328–337.

[36] A.-A. Ali, S. Wei, and L. Qian, "Optimal call admission and preemption control for public safety communications," in *IEEE Annual Conference on Information Sciences and Systems (CISS)*, 2015.

[37] ——, "Optimal admission and preemption control in finite-source loss systems," *Operations Research Letters*, vol. 43, no. 3, pp. 241–246, 2015.

[38] D. S. Sharp, N. Cackov, N. Laskovic, and L. Trajkovic, "Analysis of public safety traffic on trunked land mobile radio systems," *IEEE J. Select. Areas Commun.*, vol. 22, no. 7, pp. 1197 – 1205, September 2004.

[39] H. H. Hoang and R. Malhame and C. Rosenberg, "Communication load and delay in multichannel land mobile systems for dispatch traffic: A queueing analysis," in *IEEE Vehicular Technology Conf.*, Denver, CO, May 1992, pp. 773 – 777.

[40] R. Stevens and M. Sinclair, "Finite-source analysis of traffic on private mobile radio systems," *Electron. Lett.*, vol. 33, pp. 1292 – 1293, July 1997.

[41] R. Cooper, *Introduction to queueing theory*, 2nd ed. Elsevier North Holland, 1981.

[42] R. Ramjee, D. Towsley, and R. Nagarajan, "On optimal call admission control in cellular networks," *Wireless Networks*, vol. 3, pp. 29–41, 1997.

[43] A. Turhan, M. Alanyali, and D. Starobinski, "Optimal admission control of secondary users in preemptive cognitive radio networks," in *IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, 2012, pp. 138 – 144.

[44] S. Kockan and D. Starobinski, "Admission control and profitability analysis in dynamic spectrum access data networks," in *International Conference on Performance Evaluation Methodologies and Tools (ValueTools)*, December 2013.

[45] Z. Zhao, S. Weber, and J. C. de Oliveira, "Admission control and preemption policy design of multi-class computer networks," in *44th IEEE Annual Conference on Information Sciences and Systems (CISS)*, 2010, pp. 1–6.

[46] M. Y. Ulukus, R. Güllü, and L. Örmeci, "Admission and termination control of a two class loss system," *Stochastic Models*, vol. 27, no. 1, pp. 2–25, 2011.

[47] A. Turhan, M. Alanyali, and D. Starobinski, "Optimal admission control in two-class preemptive loss systems," *Operations Research Letters*, vol. 40, no. 6, pp. 510 – 515, 2012.

[48] S. Tozlu, M. Senel, W. Mao, and A. Keshavarzian, "Wi-fi enabled sensors for internet of things: A practical approach," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 134–143, 2012.

[49] S. Ross, *Applied Probability Models With Optimization Applications*. Dover Publications Incorporated, 1970.

[50] J. Walrand, *An Introduction to Queueing Networks*. Prentice Hall, 1988.

[51] S. A. Lippman, "Applying a new device in the optimization of exponential queuing systems," *Operations Research*, vol. 23, no. 4, pp. 687–710, 1975.

[52] S. Ross, *Introduction to Stochastic Dynamic Programming*. Academic Press, 1983.

[53] D. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Athena Scientific, 2007.

[54] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons, 2011.

[55] A.-A. Ali and S. Wei, "Modeling of coupled collision and congestion in finite source wireless access systems," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2015.

[56] ——, "Traffic-aware joint allocation of uplink control and communication channels in multicast systems," *IEEE Transactions on Vehicular Technology*, 1st revision shall be submitted in June 2015.

[57] R. Stevens and M. Sinclair, "Finite-source analysis of traffic on private mobile radio systems," *Electronics letters*, vol. 33, no. 15, pp. 1292–1293, 1997.

[58] E. Sengul, H. Koymen, and Y. Z. Ider, "A spectrally efficient PMR system utilizing broadcast service," *IEEE Transactions on Broadcasting*, vol. 51, no. 4, pp. 493–503, 2005.

[59] "FirstNet," http://www.firstnet.gov/, accessed: October, 2014.

[60] "LTE technology: The future of public safety communications," *Presentation at Public Safety Broadband Summit*, May 2013. ([online] Available at: http://broadbandsummit.apcointl.org/2013-presentations/).

[61] T. Halonen, J. Romero, and J. Melero, *GSM, GPRS and EDGE performance: evolution towards 3G/UMTS*. John Wiley & Sons, 2004.

[62] B. Furht and S. A. Ahson, *Long Term Evolution: 3GPP LTE radio and cellular technology*. CRC Press, 2009.

[63] K. A. Hafeez, L. Zhao, J. W. Mark, X. Shen, and Z. Niu, "Distributed multichannel and mobility aware cluster-based MAC protocol for vehicular ad-hoc networks (VANETs)," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 8, pp. 3886–3902, 2013.

[64] "Private communication with engineers of Louisiana Wireless Information Network (LWIN)."

[65] D. P. Bertsekas and R. G. Gallager, *Data networks*, 2nd ed. Prentice Hall, 1992.

[66] K. Kang, Y. Cho, J. Cho, and H. Shin, "Scheduling scalable multimedia streams for 3G cellular broadcast and multicast services," *Vehicular Technology, IEEE Transactions on*, vol. 56, no. 5, pp. 2655–2672, 2007.

[67] D. T. Ngo, C. Tellambura, and H. H. Nguyen, "Efficient resource allocation for OFDMA multicast systems with spectrum-sharing control," *Vehicular Technology, IEEE Transactions on*, vol. 58, no. 9, pp. 4878–4889, 2009.

[68] ——, "Resource allocation for OFDMA-based cognitive radio multicast networks with primary user activity consideration," *Vehicular Technology, IEEE Transactions on*, vol. 59, no. 4, pp. 1668–1679, 2010.

[69] H. Won, H. Cai, K. Guo, A. Netravali, I. Rhee, K. Sabnani, *et al.*, "Multicast scheduling in cellular data networks," *Wireless Communications, IEEE Transactions on*, vol. 8, no. 9, pp. 4540–4549, 2009.

[70] J. Xu, S.-J. Lee, W.-S. Kang, and J.-S. Seo, "Adaptive resource allocation for MIMO-OFDM based wireless multicast systems," *Broadcasting, IEEE Transactions on*, vol. 56, no. 1, pp. 98–102, 2010.

[71] G. Araniti, M. Condoluci, L. Militano, and A. Iera, "Adaptive resource allocation to multicast services in LTE systems," *Broadcasting, IEEE Transactions on*, vol. 59, no. 4, pp. 658–664, 2013.

[72] Y. Sun, C. E. Koksal, K.-H. Kim, and N. B. Shroff, "Scheduling of multicast and unicast services under limited feedback by using rateless codes," *arXiv preprint arXiv:1404.6683*, 2014.

[73] L. Kleinrock and F. A. Tobagi, "Packet switching in radio channels: Part i–carrier sense multiple-access modes and their throughput-delay characteristics," *Communications, IEEE Transactions on*, vol. 23, no. 12, pp. 1400–1416, 1975.

[74] S. Okasaka, "Control channel traffic design in a high-capacity land mobile telephone system," *IEEE Transactions on Vehicular Technology*, vol. 27, no. 4, pp. 224–231, 1978.

[75] D. Kuypers and P. Sievering, "Performance analysis of the random access protocol in TETRAPOL trunked radio networks," Berlin, Sep 2001. [Online]. Available: http://www.comnets.rwth-aachen.de/KuSi_MMB01.6360.0.html

[76] R. Schoenen, R. Halfmann, and B. H. Walke, "MAC performance of a 3GPP-LTE multihop cellular network," in *IEEE International Conference on Communications (ICC)*, 2008, pp. 4819–4824.

[77] J. Mo, H.-S. So, and J. Walrand, "Comparison of multichannel MAC protocols," *Mobile Computing, IEEE Transactions on*, vol. 7, no. 1, pp. 50–65, Jan 2008.

[78] C.-Y. Chang, C.-T. Chang, T.-C. Wang, and Y.-J. Lu, "STB-MAC: Staggered multichannel traffic balanced MAC protocol in wireless networks," *Vehicular Technology, IEEE Transactions on*, vol. 63, no. 4, pp. 1779–1789, May 2014.

[79] Z. Zhou, Z. Peng, J.-H. Cui, and Z. Jiang, "Handling triple hidden terminal problems for multichannel MAC in long-delay underwater sensor networks," *Mobile Computing, IEEE Transactions on*, vol. 11, no. 1, pp. 139–154, Jan 2012.

[80] Y. Peksen, C. Yaman, T. Acarman, and A. Peker, "Multi-channel operation and GNSS correction issues," in *ITS Telecommunications (ITST), 2013 13th International Conference on*, Nov 2013, pp. 468–473.

[81] Q. Wang, S. Leng, H. Fu, and Y. Zhang, "An IEEE 802.11p-based multichannel MAC scheme with channel coordination for vehicular ad hoc networks," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, no. 2, pp. 449–458, June 2012.

[82] M. H. Ahmed, "Call admission control in wireless networks: A comprehensive survey." *IEEE communications Surveys and Tutorials*, vol. 7, no. 1-4, pp. 50–69, 2005.

[83] M. Z. Chowdhury, Y. M. Jang, and Z. J. Haas, "Call admission control based on adaptive bandwidth allocation for wireless networks," *Communications and Networks, Journal of*, vol. 15, no. 1, pp. 15–24, 2013.

[84] Y. Kim, H. Ko, S. Pack, W. Lee, and X. Shen, "Mobility-aware call admission control algorithm with handoff queue in mobile hotspots," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 8, pp. 3903–3912, 2013.

[85] F. Zarai, K. B. Ali, M. S. Obaidat, and L. Kamoun, "Adaptive call admission control in 3GPP LTE networks," *International Journal of Communication Systems*, vol. 27, no. 10, pp. 1522–1534, 2014.

[86] A. Srivastava, S. Srivastava, and K. Mitchell, "Determining traffic and control channels for a packet based cellular system having time-varying traffic," in *IEEE Vehicular Technology Conference*, May 2005.

[87] C. Bordenave, D. McDonald, and A. Proutiere, "Performance of random medium access control, an asymptotic approach," *ACM SIGMETRICS Performance Evaluation Review*, vol. 36, no. 1, pp. 1–12, 2008.

[88] ——, "Asymptotic stability region of slotted Aloha," *Information Theory, IEEE Transactions on*, vol. 58, no. 9, pp. 5841–5855, 2012.

[89] R. Garcés and J. Garcia-Luna-Aceves, "A near-optimum channel access protocol based on incremental collision resolution and distributed transmission queues," in *IEEE INFOCOM*, 1998.

[90] E. Ziouva and T. Antonakopoulos, "CSMA/CA performance under high traffic conditions: throughput and delay analysis," *Computer communications*, vol. 25, no. 3, pp. 313–321, 2002.

[91] A. Pantelidou and A. Ephremides, "Wireless multicast optimization: A cross-layer approach," *Information Theory, IEEE Transactions on*, vol. 57, no. 7, pp. 4333–4343, 2011.

[92] H. V. Poor, *An introduction to signal detection and estimation*. Springer, 1994.

[93] "Advancing 3GPP networks: Optimisation and overload management techniques to support smart phones," *GSM Association (GSMA) White Paper*, 2012.

[94] A.-A. Ali, S. Wei, and J. Zhang, "Context-tree based segmentation of non-stationary sequences using a revised Bayesian information criterion," manuscript is under preparation and shall be submitted for a conference in June 2015.

[95] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1007–1016, 2006.

[96] R. Gwadera, A. Gionis, and H. Mannila, "Optimal segmentation using tree models," *Knowledge and Information Systems*, vol. 15, no. 3, pp. 259–283, 2008.

[97] "Open data BR," ([online] Available at: https://data.brla.gov/, accessed on March 2015).

Vita

Ahsan-Abbas Ali received his BSc in Electrical Engineering from the University of Engineering and Technology (UET), Lahore, Pakistan in 2007. From 2007 to 2009, he worked as a Research Associate at the Networks and Communications Lab at the Lahore University of Management Sciences, Pakistan. He also served as a Lecturer/Lab-Engineer at UET Lahore from 2008 to 2010. He is currently a PhD student at the school of Electrical Engineering and Computer Science (EECS) at the Louisiana State University (LSU), Baton Rouge. He has also been working at the same school as a graduate research and teaching assistant, since 2011. He has been awarded with the LSU Graduate Tuition Award 2011 and the ECE James R. Lewis Scholarship 2014. His research interests include modeling and analysis of communication systems, and techniques for applied data mining.